Some Results Concerning Off-Training-Set and IID Error for the Gibbs and the Bayes Optimal Generalizers

David H. Wolpert
The Santa Fe Institute
1399 Hyde Park Rd.
Santa Fe, NM 87501
email: dhw@santafe.edu

Emanuel Knill
CIC-3 Computer Research and Applications, MSB265
LANL, Los Alamos, NM 87545
email: knill@lanl.gov

Tal Grossman
Theoretical Division and CNLS, MS B213
LANL, Los Alamos, NM 87545
email: tal@goshawk.lanl.gov

December 10, 1997

Abstract

In this paper we analyze the average behavior of the Bayes-optimal and Gibbs learning algorithms. We do this both for off-training-set error and conventional IID error (for which test sets overlap with training sets). For the IID case we provide a major extension to one of the better known results of [7]. We also show that expected IID test set error is a non-increasing function of training set size for either algorithm. On the other hand, as we show, the expected off training-set error for both learning algorithms can *increase* with training set size, for non-uniform sampling distributions. We characterize what relationship the sampling distribution must have with the prior for such an increase. We show in particular that for uniform sampling distributions

and either algorithm, the expected off-training set error is a non-increasing function of training set size. For uniform sampling distributions, we also characterize the priors for which the expected error of the Bayes-optimal algorithm stays constant. In addition we show that for the Bayes-optimal algorithm, expected off-training-set error can increase with training set size when the target function is fixed, but if and only if the expected error averaged over all targets decreases with training set size. Our results hold for arbitrary noise and arbitrary loss functions.

1 Introduction

In the supervised learning problem there is an unknown *target* relationship f between an input space and an output space. A *training set* of input-output pairs is generated from the target relationship, often by directly sampling it. The problem is to use the training set to guess the input-output relationship which "best fits" the target relationship according to some suitable cost function. Such a guessed relationship from input to outputs is known as a *hypothesis* relationship. An algorithm which produces a hypothesis relationship from a training set is called a *generalizer* or a *learning algorithm*.

Conventionally, a learning algorithm's performance is measured by the "IID" (independent identically distributed) error, which uses test sets produced by the same process that generated the training set. IID error allows overlap between training and test sets. If one instead concentrates on off training-set (OTS) error, a set of "no-free-lunch" (nfl) theorems apply: averaged over all targets or averaged over all priors, all generalizers perform the same [22].

The nfl theorems do not address the issue of how well a generalizer performs if it is coupled to the prior. This paper is an investigation of this issue. We analyze several different aspects of the OTS behavior of Bayes-optimal and Gibbs learning algorithms [7, 12, 16, 22] in the case where the algorithms' assumption for the prior is correct. Although we concentrate on OTS behavior, we also analyze some aspects of the IID behavior of those algorithms.

Here we consider average, rather than worst case results for Bayes-optimal and Gibbs algorithms. There is a large body of literature concerning such average case behavior for Bayes-optimal algorithms in very general contexts [4]. This paper extends the work on the supervised learning version of this issue, for the Gibbs and Bayes-optimal algorithms.

The previous work most closely related to ours is that of Cussens [6] and of Haussler et al. [7]. Like our work, the work in [6] considers OTS behavior of Bayes-optimal algorithms. However the issues it addresses differs from those addressed in this paper. (Cussens is primarily concerned with identifying priors for which a particular learning algorithm is Bayes-optimal.) In contrast, the issues addressed by Haussler et al. are broadly similar to those addressed in this paper. The following differences between our work and that of [7] illustrate some of these issues.

The most important difference between the work in this paper and that in [7] is that we concentrate on *learning curves*, averaging over both the input and output components of the training set, and we do so for both IID and OTS error. In contrast, the work in [7] deals either with the case where the training set and test set inputs are fixed (while the outputs can vary), or, when those inputs are free to vary, considers only IID error.

Another important difference is that Haussler et al. make the restrictions that there is

no noise, the output space is binary, and the loss function is the zero-one loss. Obviously the precise forms of their bounds change when these restrictions are relaxed. Unfortunately though, there are also other changes that are far more important than the precise values of pathological-case bounds. For example, their work indicates that the "Bayes-optimal" algorithm may perform worse than the Gibbs algorithm when the prior is mis-specified. However this result does not extend from their scenario to the general case. In particular, for quadratic loss and real-valued outputs, the Gibbs algorithm is always inferior, even for a mis-specified prior; for such a scenario, guessing the average of a stochastic algorithm always beats that algorithm [23].

In addition, some of their results that one would expect to fail when their restrictions are relaxed do not. For example, perhaps the best-known of their results is that in their scenario, for a correctly-specified prior, the Gibbs algorithm can do no worse than twice as poorly as the Bayes-optimal algorithm. One might expect that the factor of 2 reflects the fact that their output space is binary. However it is straight-forward to show that this result holds for any countable output space, even if there is noise, and even if one doesn't average over output components of data sets (as Haussler et al. do); it relies only on the fact that zero-one loss is used.[1]

In contrast to all this, the primary results in our paper are explicitly shown to hold for arbitrary noise, arbitrary output spaces, and (almost) arbitrary loss functions. Moreover, restrictions on these quantities that are imposed for some of our results are quite weak. For example, one of our primary results imposes a restriction on the loss function that is met by the kinds of loss functions (e.g., L^p, zero-one) commonly encountered in the literature.

On the other hand, there are instances in which we illustrate some of our primary results with examples that are special cases. Being examples, those instances are based on assumptions; in particular, all of these secondary results rely on assumptions concerning the sampling distribution. Since Haussler et al. fix the input components of the training set and the test set in all of their analysis, the sampling distribution is irrelevant for most of their results.

As mentioned above, Haussler et al. also consider the scenario where the learning algorithm's assumption for the prior is incorrect. This scenario is beyond the scope of this paper. The interested reader should be aware that in addition to the work in [7], the general problem of incorrect priors in Bayesian analysis has also been studied in the statistics community. See [8, 1, 11]

In Section 2, the mathematical framework we will use is described. We then give a summary of why OTS error is of interest, briefly review the nfl theorems, and define the

[1] See the discussion below of Bayes-optimal and Gibbs algorithms.

Bayes-optimal and Gibbs generalizers formally.

In Section 3 we present our results. First we derive the central theorems relating the prior, the sampling distribution, and the kind of error (OTS, IID, etc.) to the learning curve. These theorems show how it is that OTS error can increase with training set size, on average, even for the Bayes-optimal algorithm (see [22].) They also show that this behavior cannot occur for IID error, regardless of the sampling distribution. (This is a stronger result than simply providing shrinking upper bounds on such error, as is done in [7].) In addition those theorems show that this behavior cannot occur for OTS error, when the input space sampling distribution is uniform. We go on in Section 3 to relate the constancy of OTS error with training set size (as opposed to decreasing OTS error) to specific properties of the prior.

Next in that section we extend the analysis to the Gibbs algorithm, and briefly explore the connections between IID and OTS error results. We also examine behavior conditioned on the number of distinct elements in the training set rather than the overall number of elements of in the training set. (Especially in scenarios of finite noise, like that studied in [7], one might consider the number of distinct elements in the training set to be a more meaningful description of that set than the total number of elements in it.) As we show, there are some somewhat counter-intuitive aspects to behavior under such conditioning.

In Section 4 we work through an example of OTS learning curves in detail, for the Bayes-optimal and Gibbs algorithms, for the "circuit prior". Finally, in Section 5 we present open issues.

2 Preliminaries

2.1 Mathematical Formalism and Notation

In this paper we use a version of the extended Bayesian formalism, much of which is described in detail in [22]. We use n and r to indicate the number of elements in $\mathcal{X}$ (the input, or instance space) and $\mathcal{Y}$ (the output space), respectively. Most of our results are not limited to finite or countable input and output spaces. However those cases are the simplest to present. Moreover, in principle they underly real world experiments where data is measured with finite precision instruments and manipulated on finite size computers.

In order to distinguish random variables from the values they can take on, we use the convention that capital roman letters indicate random variables, and corresponding lower case letters indicate a particular value of that random variable. For example the random variable F takes on the value f.

We use P to indicate both probabilities and probability density functions. Almost always we will be able to use the argument of the P to indicate both the associated random variable

and the value it takes on. For example $P(f)$ means the prior probability over the random variable F, evaluated at the value f. For those rare occasions when more precision is needed, subscripts on the letter P will indicate the associated random variable, and the argument of P will indicate the value at which it is being evaluated. For example, $P_F(f) = P(f)$. (In contrast, $P_F(h)$ means the prior probability over the random variable F, evaluated at the value h.) Finally, when such extra precision is required in describing a conditional probability, we will put the conditioning bar in the subscript so that it is clear how arguments line up with random variables. So for example, $P_{D|F}(d \mid h)$ means $\frac{P_{D,F}(d,h)}{P_F(h)}$. Finally, we use E to indicate expectation values. So for example $E(S \mid t) = \int dsP(s \mid t)\ s = \int dsP_{S|T}(s \mid t)\ s$. (Note that all variables not explicitly listed in such an expectation are implicitly averaged over.) This is common statistics notation [14].

Our primary random variables will be target relationships F, hypothesis relationships H, training sets D, and *cost* or *error* values C. Here it will suffice to have instances of the target (f) and the hypothesis (h) relationships be $\mathcal{X}$-conditioned distributions over $\mathcal{Y}$ values. For current purposes, testing (involved in determining the value of C) consists of comparing the hypothesis with the target at some $\mathcal{X}$ value. Accordingly, we will find it useful to relate our four primary random variables to one another using three other random variables.

We denote the *test point* $\mathcal{X}$ value at which testing is done by the random variable Q. We will need to distinguish the $\mathcal{Y}$ values associated with sampling the target at such a test point from the $\mathcal{Y}$ values associated with sampling the hypothesis there. We do this by introducing the random variables Y_F and Y_H, both of which are $\mathcal{Y}$-valued variables. Formally, Y_F and Y_H are defined by

$$P_{Y_F|F,Q,D,Y_H,H}(y_F \mid f, q, d, y_H, h) = P_{Y_F|F,Q}(y_F \mid f, q) \equiv P(y_F \mid f, q) \equiv f(q, y_F) \ ,$$

and similarly

$$P_{Y_H|H,Q}(y_H \mid h, q) \equiv P(y_H \mid h, q) \equiv h(q, y_H) \ .$$

In genearl, C is a function of Y_F and Y_H. If for each q, $h(q, y_H)$ is a delta function distribution over y_H (i.e., if it specifies a single-valued function from $\mathcal{X}$ to $\mathcal{Y}$), h is called *single-valued*, and similarly for f.

The value d of the training set random variable is an ordered set of m input-output examples.[2] Those examples are indicated by $\{(d_\mathcal{X}(i), d_\mathcal{Y}(i))\}_{i=1..m}$. The ordered set of all input values in d is $d_\mathcal{X}$ and similarly for $d_\mathcal{Y}$. The number of distinct values in $d_\mathcal{X}$ is denoted by m'.

The *likelihood* specifies how the data is generated. In the case of *IID* (independent identically distributed) generation of d, d is formed by sampling $\mathcal{X}$ according to a *sampling*

[2]We apologize for using the now-traditional term "training *set*" to refer to what is properly a sequence.

distribution $\pi(x)$ and then sampling f at those points in $\mathcal{X}$. More formally, in this paper we assume the IID likelihood given by

$$P(d \mid f) = \prod_{i=1}^{m} \pi\left(d_{\mathcal{X}}(i)\right) f\left(d_{\mathcal{X}}(i), d_{\mathcal{Y}}(i)\right). \tag{1}$$

The *posterior*, $P(f \mid d)$, is the Bayesian inverse of the likelihood.

As an example, in the noise-free IID case, f is single-valued and $P(d \mid f)$ is given by $\delta(d \subseteq f) \prod_i \pi(d_{\mathcal{X}}(i))$, where $\delta(d \subseteq f) = 1$ if d agrees with f, and 0 otherwise. Bayes' theorem then gives us $P(f \mid d) \propto P(f)\ \ \delta(d \subseteq f) \prod_i \pi(d_{\mathcal{X}}(i))$.

It is important to note that in this paper we are *not* assuming a noise free situation — our results hold for arbitrary f, as long as the likelihood is IID. This means in particular that for the scenarios considered in this paper, two elements in the training set can have the same $\mathcal{X}$ values but different $\mathcal{Y}$ values.

Any (!) learning algorithm is simply a distribution $P(h \mid d)$. It is *deterministic* if the same d always gives the same h (i.e., if for fixed d, $P(h \mid d)$ is a delta function about one particular h). A non-deterministic learning algorithm is known as *stochastic*. We will say that a full joint distribution $P(h, d, f, c, y_F, y_H, q)$, "has" or "specifies" the generalizer given by

$$P(h \mid d) = \frac{\int df \sum_{c, y_F, y_H, q} P(h, d, f, c, y_F, y_H, q)}{\int df\, dh' \sum_{c, y_F, y_H, q} P(h', d, f, c, y_F, y_H, q)}.$$

In supervised learning $P(h \mid f, d) = P(h \mid d)$ (i.e., the learning algorithm only sees d in making its guess, not f). Similarly, $P(f \mid h, d) = P(f \mid d)$, and therefore the distribution $P(h, f \mid d) = P(h \mid d) P(f \mid d)$ so that H and F are conditionally independent given D.

With some similar stipulations, it follows that (for example) $P(y_F \mid y_H, d, q) = P(y_F \mid d, q)$. (To see this, expand both sides in terms of f; use Bayes' theorem to interchange f and y_H; then expand both distributions over y_H in terms of h.) It also follows that

$$P(h, d, f, c, y_F, y_H, q) = P(h \mid d)\, P(d \mid f)\, P(q \mid d)\, P(f)\, h(y_H, q)\, f(y_F, q)\, P(c \mid y_F, y_H).$$

It does *not* follow that $P(y_F \mid y_H, d) = P(y_F \mid d)$. Intuitively, this is because for a fixed learning algorithm, knowing y_H and d can tell you something about q that knowing d alone does not. Therefore knowing y_H (in conjunction with knowing d) can tell you something about y_F that knowing d alone does not. The issue of when you can (not) dispense with a y_H in the conditioning event are crucial for the precise formulation of our results.

We define *IID error* to mean $P(q \mid d) = \pi(q)$. Similarly we define *OTS error* by the OTS sampling distribution

$$P(q \mid d) = \frac{\pi(q)\delta(q \notin d_{\mathcal{X}})}{\sum_q \pi(q)\delta(q \notin d_{\mathcal{X}})} \tag{2}$$

where $\delta(z) = 1$ if z is true, 0 otherwise. (When the argument of a $\delta(.)$ is a numeric rather than logical value, it is implicitly a Dirac delta function rather than an indicator function.) Our analysis up to Theorem 3.2 holds for either IID or OTS error. After that, unless explicitly stated otherwise, we will implicitly restrict attention to OTS error.

The random variable cost C is defined by $C = L(Y_H, Y_F)$, where $L(.,.)$ is called the *loss function*. For example, zero-one loss has $L(a,b) = 1 - \delta(a = b)$. As another example, quadratic loss has $L(a,b) = (a-b)^2$.

To relate analysis involving the random variable C to other frameworks that may be more familiar to the reader, first write

$$E(C \mid f, h, d) = \sum_{y_H, y_F, q} E(C \mid f, h, d, y_H, y_F, q) P(y_H, y_F, q \mid f, h, d).$$

Due to our definition of C, the first term in the sum equals $L(y_H, y_F)$. The second term equals $P(y_H \mid h, q, f, d, y_F) P(y_F \mid q, f, d, h) P(q \mid d, f, h)$. This in turn equals $h(q, y_H) f(q, y_F) P(q \mid f, h, d)$. In this paper, we will always have $P(q \mid f, h, d) = P(q \mid d_{\mathcal{X}})$. Therefore

$$\bar{C} \equiv E(C \mid f, h, d) = \sum_{y_H, y_F, q} L(y_H, y_F) h(q, y_H) f(q, y_F) P(q \mid d_{\mathcal{X}}). \tag{3}$$

In many treatments of supervised learning (e.g., much of computational learning theory) one analyzes the behavior of the *generalization error* $\bar{C}$. In particular, if not all three of the values f, h and d are known, then the value of $\bar{C}$ is not fixed, and probability distributions over $\bar{C}$ can be analyzed. This is done in PAC for example [22].

In practice, the real-world problem at hand should determine whether one is interested in C or $\bar{C}$. Note that distributions over C do not fix the associated distribution over $\bar{C}$ in general, nor vice-versa. This despite the fact that $\bar{C}$ is simply an average of C.[3]

So for example PAC results concerning $\bar{C}$ do not immediately translate into PAC results concerning C. However expectations of C that are not conditioned on q, y_H or y_F are identical to the corresponding expectation of $\bar{C}$ (e.g., $E(C \mid d) = E(\bar{C} \mid d)$). Since in this

[3] As a simple example of this, consider the case where $r = 2$ and $|\mathcal{X}| = m' + 2$. Furthermore, say that for our given d and the resultant h, either f agrees exactly with h for all off-training set q or the two never agree, with equal probability. (I.e., only those two kinds of f have non-zero posterior probability, and the posterior probability for each kind is the same.) This means that for OTS error and the zero-one loss function, $P(c \mid d) = \delta(c = 0)/2 + \delta(c = 1)/2$, which here equals $P_{\bar{C}|D}(c \mid d)$. However $P(c \mid d)$ would have the same form if all four possible relationships between h and f for the off-training set q had non-zero posterior probability. And in that second case, we would have the possibility of $\bar{C}$ values that are impossible in the first case. See [23].

paper we are only concerned with such expectation values, without loss of generality we can restrict attention to C and all of our results still hold for $\bar{C}$.

2.2 OTS Error and the NFL Theorems

Many supervised learning texts take the view that "the overall objective ... is to learn from samples and to generalize to new, as yet unseen cases" [Weiss and Kulikowski 1991]. Similarly, it is common practice to try to avoid fitting the training set exactly when one learns from that set, i.e., to try to avoid "overtraining". One of the major rationales given for this is that if one overtrains, "the resulting [system] is unlikely to classify additional points [in the input space] correctly" [5]. As another example, in [3], we read that "the real value of a scientific explanation lies not in its ability to explain [what one has already seen], but in predicting events that have yet to (be seen)". As a final example, in [10] we read that "[in Machine Learning we wish to know whether] any hypothesis found to approximate the target function well over a sufficiently large set of training examples will also approximate the target function well over other unobserved examples."

This language makes clear that OTS behavior is a central concern of supervised learning, even though little theoretical work has been devoted to it to date. Some of the reasons for such concern are as follows.

1. In the low-noise regime, optimal behavior on the training set is equivalent to look-up table memorization, and the only interesting issues concern OTS behavior. In particular, in that low-noise regime, if one uses a memorizing learning algorithm, then for test sets overlapping with training sets the upper limit of possible test set error values shrinks as the training set grows. If one doesn't correct for this when comparing behavior for different sizes of the training set (as when investigating learning curves), one is comparing apples and oranges. In that low-noise regime, correcting for this effect by renormalizing the range of possible errors is equivalent to requiring that test sets and training sets be distinct. (See [22].)

2. Very often the process generating the training set is not the same as that governing testing. In such scenarios, the usual justification for testing with the same process that generated the training set (and with it the possibility that test sets overlap with training sets) doesn't apply.

   One example of such a difference between testing and training is *active* or *query-based* or *membership-based* learning. In that kind of learning the learner chooses, perhaps dynamically, where in the input space the training set elements are to be. However,

conventionally, there is no such control over the test set. So testing and training are governed by different processes.

As another example, say we wish to learn tertiary protein structure from primary structure and then use that to aid drug design. We already know what tertiary structure corresponds to the primary structures in the training set. So we will never have those sequence structure pairs in the *test set* (i.e., in the set of amino acid sequences whose tertiary structure we wish to infer to aid the drug design process). We are therefore very interested in OTS error.

3. Since behavior can be radically different in the regime where test examples coincide with the training set from the behavior where there is no overlap, it is natural to distinguish them.

4. When the training set is much smaller than the full input space, the probability that a randomly chosen test set input value coincides with the training set is vanishingly small (assuming a non-pathological sampling distribution). In such situations one expects the value of the OTS error to be well-approximated by the value of the conventional IID error, an error which allows overlap between test sets and training sets.

   One might suppose that in such a small training set regime there is no aspect of OTS error not addressable by instead calculating IID error. This is wrong, as the following several points illustrate.

   (a) First, even if OTS error is well approximated by IID error, it does not follow that quantities derived from the errors are close to one another. For example, it does not follow that the sign of the slope of the learning curve - often an object of major interest - is the same for both errors over some region of interest.

   (b) Second, although it's usually true that a probability distribution *over* IID error will well-approximate the corresponding distribution over OTS error, distributions *conditioned on* IID error can differ drastically from distributions conditioned on OTS error.

   As an example, let s be the empirical misclassification rate between a hypothesis and the target over the training set, $\bar{c}_{IID}$ the misclassification rate over all of the input space (the IID zero-one generalization error), and $\bar{c}_{OTS}$ the misclassification rate over that part of the input space lying outside of the training set. Assume a uniform distribution over the input space, a uniform prior over target input-output relationships, and a noise-free IID likelihood (Eq.1). Then $P(s \mid \bar{c}_{IID}, m)$ is just $(\bar{c}_{IID})^{sm} \quad (1 - \bar{c}_{IID})^{(m-sm)} \quad \binom{m}{sm}$ (s can be viewed as the percentage of

heads in m flips of a coin with bias $\bar{c}_{IID}$ towards heads). On the other hand, $P(s \mid \bar{c}_{OTS}, m)$ is independent of $\bar{c}_{OTS}$. (This is proven in [23].)

(c) Third, often it is more straight-forward to calculate a certain quantity for OTS rather than IID error. In such cases, even if one's ultimate interest is IID error, it makes sense to instead calculate OTS error (assuming OTS error well-approximates IID error).

As an example, OTS error results presented in [23] mean that when the training set is much smaller than the full input space, $P(\bar{c}_{IID} \mid s, m)$ is (arbitrarily close to) independent of s, if the prior over target input-output relationships is uniform. (This holds despite VC results that it is highly unlikely for $\bar{c}_{IID}$ and s to differ significantly, no matter what the prior [18]. The resolution of this apparent contradiction ia not difficult, and can be found in [22].)

As another example, one of the more quoted results from [7] is that for a mis-specified prior, the Gibbs algorithm may perform better than the Bayes-optimal algorithm. However by one of the no-free-lunch theorems [23], averaged over all priors, the expected OTS error of any two fixed learning algorithms is identical. So for large enough input spaces, we immediately see that there must be (mis-specified) priors for which the Gibbs algorithm beats the Bayes-optimal algorithm. There is no need to invoke the information-gain machinery developed in [7] to get this result. Indeed, by the no-free-lunch theorems, we immediately further see that the probability mass of such priors where the Gibbs algorithm wins must exactly equal that of priors for which the Bayes-optimal algorithm wins. (See [23].)

None of this means that the conventional error measure is "wrong". No claim is being made that one should not test with the same process that generated the training set, to avoid having error measure accuracy on the training set as well as off it. Rather the claim is simply that OTS testing is an issue of major importance. In that it gives no credit for memorization, it is also the natural way to investigate whether one can make assumption-free statements concerning an algorithm's *generalization* ability.

Unfortunately, for OTS error we have the afore-mentioned "no-free-lunch theorems" limiting the assumption-free utility of any learning algorithm. In particular, consider the uniform average, over all prior distributions $P(f)$, of the expected cost given m, $E(C \mid m)$. This is given by the average of $\sum_{f,h,d} E(C \mid f, h, d)\ P(f, h, d \mid m)$. One of the no-free-lunch theorems says that for the zero-one loss function, this is the same for *all* fixed learning algorithms (i.e., for all algorithms that do not vary as $P(f)$ changes). Loosely speaking, for OTS error, there are "just as many" priors in which any algorithm A has superior behavior to that of any

other algorithm B as vice-versa. Similar results hold for other loss functions, and for other conditioning events (besides m). In addition, due to the relationship between IID and OTS error discussed in point 4 above, these results carry over to IID error in the appropriate limit. (See [23].)

Of course, a natural way to extend these theorems is to consider the case where the uniformity of the average over priors is relaxed. Unfortunately, this alone could just as easily hurt one's learning algorithm as help it. However if the learning algorithm "knows" the prior, presumably it should be able to use this information to ensure that it performs well. In this paper we investigate the case where the learning algorithm and $P(f)$ are coupled.

2.3 The Gibbs and Bayes-optimal generalizers

The Bayes-optimal generalizer is the one that minimizes expected cost given d, $E(C \mid d)$. In other words it produces the best (as far as expectation value is concerned) guess one could, given the training set at hand. It is deterministic, with its hypothesis $h^\star$ given by

$$h^\star(x) = \text{argmin}_y W(x, y) \tag{4}$$

where $W(x, y) \equiv \int df L(f(x), y) P(f \mid d)$. See [7, 12, 22]. (In cases of multiple minima of $W(x, y)$, for current purposes any tie-breaking scheme will do.)

For the zero-one loss and single-valued f (i.e., $P(f)$ that equal 0 for non-single-valued f), $h^\star(x) = \text{argmax}_y \sum_f \delta(f(x) = y) P(f \mid d)$. In particular, if $P(f)$ is uniform across all single-valued f in some *target class* U and zero otherwise, and if $\mathcal{Y} = \{0, 1\}$, then for any x, $h^\star(x) = 1$ if the number of f in U that are consistent with d and go through the point $(x, 1)$ exceeds the number of f in U that are consistent with d and go through $(x, 0)$. It equals 0 otherwise.

A Gibbs generalizer is one that obeys

$$P(h \mid d) \propto P_{D|F}(d \mid h) G(h) \tag{5}$$

where $G(h)$ is a probability distribution and the proportionality constant (sometimes called a *partition function*) is set by normalization and depends only on d [7, 12, 16, 22].

A *correct* Gibbs generalizer is a Gibbs generalizer for which $G(h) = P_F(h)$. By Bayes' theorem, such a generalizer can be viewed as creating hypotheses by sampling the posterior distribution over targets. In our results below, unless explicitly stated otherwise, we restrict attention to correct Gibbs generalizers.

Intuitively, a Gibbs generalizer is one that knows the likelihood, and guesses by randomly sampling what it thinks (!) is the posterior. (If $G(h) \neq P_F(h)$, that presumption is incorrect.) When the likelihood is noise-free, such a generalizer reduces to the exhaustive learning

generalizer ([21, 15, 17]). Intuitively, that generalizer consists of the following rule: Given d, find all h which go through d; then choose randomly among those h, where "randomly" means according to the distribution $G(h)$.

In general, $G(h)$ need not equal $P(h)$. Also, much of the analysis of Gibbs generalizers is not predicated on their being correct, so that $G(h) \neq P_F(h)$. In this, it is wrong to refer to $G(h)$ as a "prior" in any sense. (Despite this, $G(h)$ is often referred to as the "assumed prior" in the literature.) In the same regard, it should be noted that the Gibbs generalizer does not in any sense "follow" from Bayes' theorem — that role is played by the Bayes-optimal generalizer [16, 7, 12, 13, 22, 21]. Indeed, in almost all situations the Bayes-optimal generalizer is deterministic, always guessing the same h for the same d. In such situations, there is no $G(h)$ for which the Bayes-optimal and Gibbs generalizers are equivalent. That is because in general there is no $G(h)$ that allows a Gibbs generalizer to be deterministic [22].

Note that in the absence of noise, the Bayes-optimal generalizer creates hypotheses that agree exactly with all the elements in the training set. The same is true of the Gibbs generalizer, if either $G(h)$ is nowhere-zero or $G(h) = P_F(h)$.

Formally speaking, whether your generalizer is Bayes-optimal is determined the entire joint distribution, $P(f, h, d, c)$. One can not say that $P(h \mid d)$ is Bayes-optimal, only that it is Bayes-optimal for a particular $P(f \mid d)$ (the posterior) and $P(c \mid f, h, d)$ (the rule for assigning cost to f, h and d). In this paper, unless explicitly stated otherwise, one can assume there is this correspondence when we refer to a Bayes-optimal generalizer. Similar comments apply to Gibbs generalizers.[4]

When $P(h \mid d)$ is the Bayes-optimal generalizer, we will write probability distributions as conditioned on BO. For example, we will write $P(h \mid d, BO)$, or $P(c \mid d, BO)$. Similarly, for Gibbs generalizers we will condition on G. Formally, this means that there is another variable, a hyperparameter specifying $P(h \mid d)$, and that two of the values of this variable are BO and G. It is implicitly assumed that specifying the value of this hyperparameter does not alter the likelihood, the posterior, etc.

To illustrate these concepts, here is the promised generalization of the result in [7] concerning the relation between the expected errors of the correct Gibbs generalizer and the Bayes optimal generalizer. Let $\max_y \sum_f P(f \mid d)\delta(f(x) = y) = b$, and let y' be the associated maximizing y. The Bayes-optimal algorithm will guess y' for that x, and thereby incur expected zero-one loss of $1 - b$. (Note the sum over y of $\sum_f P(f \mid d)\delta(f(x) = y) = 1$.) The error of the Gibbs algorithm however will be $b(1 - b)$ (the probability of guessing y' times the associated expected loss) plus a term bounded above by $1 - b$. Therefore the ratio of

[4] Though to have your generalizer be a Gibbs generalizer, the only property of the full joint distribution besides $P(h \mid d)$ that is important is $P(d \mid f)$; the loss function is irrelevant.

expected losses is bounded above by $1 + b$. QED.

It is trivial to see that one can fix f and/or d in such a way that the Bayes-optimal generalizer's guess gets worse as more elements are added to the training set. However it turns out that there exist non-uniform sampling distributions for which, for all m', $E(C \mid m', BO)$ is an increasing function of m'. In other words, *average* (over training sets and targets) OTS error can increase with the training set size for the Bayes-optimal generalizer. On average, the more you know the worse your best-possible (!) predictions on things you don't know.

This result can be established by the following example: Let X consist of three values, 0, 1 and 2, and Y consist of two values $\{0, 1\}$. Let $P(f) = 0$ for all f aside from those in some "target class" U. Assume that $P(f)$ is uniform over all f in U. Let U be the set of *all* functions f such that $f(x \in \{0, 1\}) = 1$, i.e., U is the set consisting of the function with successive output values given by (1, 1, 1) along with the function with successive output values give by (1, 1, 0). Let $\pi(x) = 1/2 - \gamma/2$ for both $x \in \{0, 1\}$. Then

$$E(C \mid m = 1, BO) - E(C \mid m = 0, BO) = \gamma[(1 - \gamma)/(1 + \gamma) - 1/2]$$

This is positive for $\gamma < 1/3$.

Similarly, there exist non-uniform sampling distributions for which, for all m', $E(C \mid m', G)$ is an increasing function of m'. This can be seen by using the same example used for the Bayes-optimal generalizer. (Note that in general, if $E(C \mid m' = 0, BO) = E(C \mid m' = 0, G)$, then since $E(C \mid m' = k, G) \geq E(C \mid m' = k, BO)$, if $E(C \mid m', BO)$ increases when m' goes from 0 to k, so must $E(C \mid m', G)$.)

As another perhaps counterintuitive example, $E(C \mid f, m)$ (i.e., the expected OTS error for fixed f) can increase with m. This can happen even when we have a Bayes-optimal generalizer, $n \gg m$, $\pi(x)$ is uniform, and f is a target such that $P(f)$ is nonzero (see appendix 1 in [22]).

In all of these instances, improvement in (conventional) IID generalization error with increased training set size is due purely to the BO algorithm's ability to guess well on the training set; no "generalization" contributes. In addition to such counter-intuitive OTS phenomena, there are a number of somewhat counter-intuitive phenomena associated even with IID error, where one gets credit for guessing well on the training set. For example, as is shown in the next section, there are scenarios where increasing m' worsens IID error, even when one averages over training sets. As another example, $E(C \mid f, m)$ can increase with m for IID error. (Note that the example at the end of appendix 1 in [22] applies just as well to IID as OTS error.)

Currently it is not known how pathological these scenarios are. It is not known, for example, what fraction of all sampling distributions and priors give increasing OTS error for Bayes-optimal generalizers. However the analysis in the following section characterizes

the relationship between the sampling distribution and prior on the one hand, and whether OTS and/or IID error is increasing on the other.

3 OTS and IID learning curves

Once we are given the likelihood, for both IID and OTS error, for both the Gibbs and Bayes-optimal generalizers, the "learning curve" $E(C \mid m)$ is governed by the interplay between the sampling distribution and the prior. To see this, let "$d(m)$" mean the m'th pair in d, "d^{m-1}" mean the first $m-1$ pairs in d, and similarly for other superscripts. So for example $d^m_{\mathcal{X}} = \{d^{m-1}_{\mathcal{X}}, d_{\mathcal{X}}(m)\}$. Then the central role of the interplay is succinctly expressed by the following equation:

$$E(C \mid m) = \sum_{d^m_{\mathcal{X}}} P(d^m_{\mathcal{X}} \mid m) \sum_q P(q \mid d^m_{\mathcal{X}}, m) E(C \mid q, d^m_{\mathcal{X}}, m) \ . \tag{6}$$

The last term on the right-hand side is determined solely by $P(f)$ (assuming a given likelihood). The other two terms in that sum are determined solely by $\pi(x)$.

However to analyze the m-dependence of learning curves, it is in many ways easier to work with formulas where the terms in the summand entangle the prior and the sampling distribution. To that end, first note that $P(q \mid d) = P(q \mid d_{\mathcal{X}})$ and $P(d^{m-1} \mid m) = P(d^{m-1} \mid m-1)$, where the following notation is being used: Conditioning on m or $m-1$ means the size of the training set is m or $m-1$, respectively, and in both cases, d^{m-1} is an ordered set of the first $m-1$ values of the training set. So for example, "$P(d^{m-1} \mid m)$" is the probability that in an m-element training set, the first $m-1$ elements are given by d^{m-1}.

Using this notation, we can write down the following "entangled" formulas, which involve expanding over all of d^{m-1} rather than $d^m_{\mathcal{X}}$:

$$E(C \mid m) = \sum_{d^{m-1}} P(d^{m-1} \mid m-1) \sum_q P(q \mid d^{m-1}_{\mathcal{X}}, m) E(C \mid q, d^{m-1}, m) \ ,$$

$$E(C \mid m-1) = \sum_{d^{m-1}} P(d^{m-1} \mid m-1) \sum_q P(q \mid d^{m-1}_{\mathcal{X}}, m-1) \ E(C \mid q, d^{m-1}, m-1) \ . \tag{7}$$

Note that $E(C \mid q, d^{m-1}, m)$, to give one example, depends on both the sampling distribution (through the un-fixed value of $d_{\mathcal{X}}(m)$) and the prior.[5]

The reason it is relatively easy to work with Eq. 7 is that, as is proven in Theorem 3.2 below, $E(C \mid q, d^{m-1}, m, BO) \leq E(C \mid q, d^{m-1}, m-1, BO)$, for all q and d^{m-1}, independent

[5] In examining Eq. 7, bear in mind that $P(q \mid d^{m-1}_{\mathcal{X}}, m) = \sum_{d_{\mathcal{X}}(m)} P(q \mid d^m_{\mathcal{X}}, m) P(d_{\mathcal{X}}(m) \mid d^{m-1}_{\mathcal{X}}, m) = \sum_{d_{\mathcal{X}}(m)} P(q \mid d^m_{\mathcal{X}}, m) P(d_{\mathcal{X}}(m))$. This need not equal $P(q \mid d^{m-1}_{\mathcal{X}}, m-1)$, for example when we have OTS error. See below.

of the form of $P(q \mid d)$. (The same holds for the Gibbs generalizer, provided some restrictions on the loss function are met.) Now in some scenarios $P(q \mid d_{\mathcal{X}}^{m-1}, m-1)$ and $P(q \mid d_{\mathcal{X}}^{m-1}, m)$ differ, whereas in others they are the same. When they are the same, our inequality and Eq. 7 mean that the learning curve is non-increasing.

However when they differ, it is possible that the learning curve can increase. This can happen if the likely test set element changes with the additional training set element from q to q', where $E(C \mid q', d^{m-1}, m) > E(C \mid q, d^{m-1}, m-1)$. Note that such an inequality does not contradict the fact that for any *fixed* test set element, this expectation value is shrinking. $E(C \mid q, d^{m-1}, m)$ is decreasing for all q, but we are "jumping" to q's at progressively higher points in the decline of their $E(C \mid q, d^{m-1}, m)$.

3.1 Central Results for Bayes-Optimal Learning Curves

To prove that $E(C \mid q, d^{m-1}, m) \leq E(C \mid q, d^{m-1}, m-1)$, we start with the following lemma which holds for the IID likelihood (and not necessarily for other likelihoods). In essence, the lemma says that for that likelihood, one can marginalize training set elements if one is careful about the training set size in the conditioning event:

Lemma 3.1 *For the IID likelihood (Eq. 1),*

$$\sum_{d(m)} P(d(m), y_F \mid q, d^{m-1}, m) = P(y_F \mid q, d^{m-1}, m-1) \ . \tag{8}$$

Proof: Perform the sum over $d(m)$ to reduce our problem to showing that

$$P(y_F \mid q, d^{m-1}, m) = P(y_F \mid q, d^{m-1}, m-1) \ \ . \tag{9}$$

This always holds, due to our likelihood. To see this, expand in terms of f, getting

$$P(y_F \mid q, d^{m-1}, m) = \int df f(q, y_F) \ P(f \mid q, d^{m-1}, m)$$

(and similarly for $P(y_F \mid q, d^{m-1}, m-1)$). Next use Bayes' theorem to write

$$P(f \mid q, d^{m-1}, m) = \frac{P(d_{\mathcal{Y}}^{m-1} \mid f, q, d_{\mathcal{X}}^{m-1}, m) \ P(f \mid q, d_{\mathcal{X}}^{m-1}, m)}{P(d_{\mathcal{Y}}^{m-1} \mid q, d_{\mathcal{X}}^{m-1}, m)} \ , \tag{10}$$

and similarly for $P(f \mid q, d^{m-1}, m-1)$.

By our likelihood, $P(d_{\mathcal{Y}}^{m-1} \mid f, q, d_{\mathcal{X}}^{m-1}, m) = P(d_{\mathcal{Y}}^{m-1} \mid f, q, d_{\mathcal{X}}^{m-1}, m-1)$ (the fact that one is going to end up with an m-element training set rather than an $m-1$ element training set doesn't affect the probability of $d_{\mathcal{Y}}^{m-1}$). In addition, it is implicitly assumed throughout this paper that $P(f \mid q, d_{\mathcal{X}}^{m-1}, m) = P(f \mid q, d_{\mathcal{X}}^{m-1}, m-1) = P(f)$. (Neither the test set

elements nor the number of elements in the training set nor their $\mathcal{X}$ components affects the probability of a particular f.) Therefore the numerators in Eq. (10) are the same whether we condition on m or $m-1$. Since the denominators in Eq. (10) are simply sums over f of the numerators, this means that $P(f \mid q, d^{m-1}, m) = P(f \mid q, d^{m-1}, m-1)$. Plugging this into our expansions for $P(y_F \mid q, d^{m-1}, m)$ and $P(y_F \mid q, d^{m-1}, m-1)$ establishes the desired equality. **QED**

This kind of expansion in which one distribution is a sum over other distributions is called a *refinement*. Intuitively, it means that the summed distribution contains more information than the other one. Such an "information inequality" is a powerful tool. In particular, the inequality in Lemma 3.1 serves as the foundation of many of the results in this paper.

To see this, write

$$E(C \mid q, d^{m-1}, m) = \sum_{y_F, y_H, d(m)} E(C \mid y_F, y_H, d^m, q, m) P(y_H \mid y_F, d^m, q, m) P(y_F, d(m) \mid q, d^{m-1}, m). \tag{11}$$

Plugging in for C, we get

$$E(C \mid q, d^{m-1}, m) = \sum_{y_F, y_H, d(m)} L(y_F, y_H) P(y_H \mid d^m, q, m) P(y_F, d(m) \mid q, d^{m-1}, m). \tag{12}$$

Now $E(C \mid m)$ is minimized by the Bayes-optimal generalizer [22]. This means that $E(C \mid q, d^{m-1}, m)$ is also minimized by that generalizer. To see this, simply recall Eq. 7, which gives $E(C \mid m)$ as a linear combination of the $E(C \mid q, d^{m-1}, m)$, where the combination coefficients are all non-negative. If that sum is known to be minimized, then each individual $E(C \mid q, d^{m-1}, m)$ in the sum must be minimized, or we could modify our learning algorithm to change one such term and thereby get a lower overall sum.

Given Eq. 12, this implies that for any q, any distribution P, and any generalizer γ,

$$\begin{aligned} & \sum_{y_F, y_H, d(m)} L(y_F, y_H)\ P(y_H \mid d^m, q, m, BO)\ P(y_F, d(m) \mid q, d^{m-1}, m) \\ \leq & \sum_{y_F, y_H, d(m)} L(y_F, y_H)\ P(y_H \mid d^m, q, m, \gamma)\ P(y_F, d(m) \mid q, d^{m-1}, m)\ . \end{aligned} \tag{13}$$

This inequality serves as the basis for why for Bayes-optimal generalizers, in many scenarios expected cost goes down as training set size rises. Intuitively, this follows from letting γ be the Bayes-optimal learning algorithm where the training set only consists of $m-1$ elements rather than m. To establish this formally we use Lemma 3.1 to derive the following:

Theorem 3.2 *For the IID likelihood,* $E(C \mid q, d^{m-1}, m, BO) \leq E(C \mid q, d^{m-1}, m-1, BO)$.

Proof: Write

$$E(C \mid q, d^{m-1}, m-1) = \sum_{y_F, y_H} L(y_F, y_H) P(y_H \mid q, d^{m-1}, m-1) P(y_F \mid q, d^{m-1}, m-1) \ ,$$

where the BO in the conditioning event is implicit.

By Lemma 3.1, we can replace $P(y_F \mid q, d^{m-1}, m-1)$ with $\sum_{d(m)} P(y_F, d(m) \mid q, d^{m-1}, m)$. Doing this in the expansion for $E(C \mid q, d^{m-1}, m-1)$ gives

$$E(C \mid q, d^{m-1}, m-1) = \sum_{y_F, y_H, d(m)} L(y_F, y_H) P(y_H \mid q, d^{m-1}, m-1) P(y_F, d(m) \mid q, d^{m-1}, m).$$

However by Bayes-optimality (Eq.13) and Eq. 12, this is greater than or equal to

$$E(C \mid q, d^{m-1}, m) = \sum_{y_F, y_H, d(m)} L(y_F, y_H) P(y_H \mid q, d^{m}, m) P(y_F, d(m) \mid q, d^{m-1}, m) \ .$$

(Take the generalizer γ to be the Bayes-optimal generalizer that only sees the first $m-1$ elements of d rather than all of d.) **QED**

In analogy to Eq. 7, write

$$E(C \mid d^{m-1}, m) = \sum_{q} P(q \mid d_{\mathcal{X}}^{m-1}, m) E(C \mid q, d^{m-1}, m) \ ,$$

and similarly for $E(C \mid d^{m-1}, m-1)$. Now for IID test set error, it is always the case that $P(q \mid d_{\mathcal{X}}^{m-1}, m) = P(q \mid d_{\mathcal{X}}^{m-1}, m-1) \propto \pi(q)$. Accordingly, by Theorem 3.2, for IID test set error $E(C \mid d^{m-1}, m, BO) \leq E(C \mid d^{m-1}, m-1, BO)$. In addition, consider OTS error with a uniform π. Now we again have $P(q \mid d_{\mathcal{X}}^{m-1}, m) = P(q \mid d_{\mathcal{X}}^{m-1}, m-1) = \pi(q)$, by symmetry. Accordingly, for OTS error and a uniform sampling distribution, we again get $E(C \mid d^{m-1}, m, BO) \leq E(C \mid d^{m-1}, m-1, BO)$.

In other words, under those conditions, expected error shrinks along the average *trajectory* of a sequence of new elements being added to the training set. More precisely, for any fixed d, if we average over the next input-output pair to add to d, then our expected error after adding that pair can not be greater than our current expected error.

Now consider Eq. 7 itself. As when comparing the quantities $E(C \mid d^{m-1}, m, BO)$ and $E(C \mid d^{m-1}, m-1, BO)$, note that $P(q \mid d_{\mathcal{X}}^{m-1}, m) = P(q \mid d_{\mathcal{X}}^{m-1}, m-1) = \pi(q)$ for either IID error or for OTS error with uniform π. Accordingly, by Theorem 3.2, for those cases, $E(C \mid m)$ is a non-increasing function of m.

Summarizing, we have the following theorem:

Theorem 3.3 *For the IID likelihood, for either IID error, or for OTS error with uniform* π*,*

i) $E(C \mid m, BO) \leq E(C \mid m-1, BO)$,
and for any d^{m-1},
ii) $E(C \mid d^{m-1}, m, BO) \leq E(C \mid d^{m-1}, m-1, BO)$.

Note that all of these results hold for any $P(d_{\mathcal{Y}} \mid d_{\mathcal{X}}, f)$ (i.e., any noise process) and any loss function L. Note also that for OTS error and nonuniform π, $P(q \mid d_{\mathcal{X}}^{m-1}, m)$ need not equal $P(q \mid d_{\mathcal{X}}^{m-1}, m-1)$. In such a situation, even though $E(C \mid q, d^{m-1}, m) \leq E(C \mid q, d^{m-1}, m-1)$ for any *single* q, the average of it giving $E(C \mid m)$ may increase with m. This is exactly what happens in the example above of increasing OTS error. Of course, by Eq.7 and Theorem 3.2, even for such a case where expected OTS error can increase when the training set size goes from $m-1$ to m, the upper bound on such error, given by $\max_{q,d^{m-1}} E(C \mid q, d^{m-1}, m)$ and $\max_{q,d^{m-1}} E(C \mid q, d^{m-1}, m-1)$, is non-increasing.

One might hope that under the conditions of Theorem 3.2, for any particular d^m expected OTS error for a Bayes-optimal learning algorithm is $\leq$ the error for the associated d^{m-1}. This would mean that $E(C \mid BO)$ shrinks along any *particular* trajectory of input-output pairs added to the training set, not just along the average such trajectory.

This is not true however[6]

So in a certain sense, our results are as strong as possible: if you fix f rather than averaging over it (see the discussion at the end of the previous section), or fix $d(m)$ rather than averaging over it, you do not necessarily get non-increasing error. This is true even if one is concerned with IID error.

3.2 Conditioning on m', the number of unique elements in the training set

An issue closely related to learning curves is how expected error behaves if one conditions on m' and then changes m' (rather than condition on m and then change m). One natural way to investigate this would be to analyze $E(C \mid m')$ as a function of m'. In general though, only if $P(m)$ —the prior probability that we will pick m elements in our training set — is known can one can investigate the behavior of $E(C \mid m')$. This follows from the fact that when only m' is fixed, m is set by $P(m \mid m') \propto P(m' \mid m)\ P(m)$. So when we condition only on m', if $P(m)$ is not specified, then in addition what the value of m is likely to be is

[6]As an example, let $P(f)$ be non-zero for only three (noise-free) f's: the f that is all 0's (probability = .9), the f that is all 1's (probability = .05), and the f that is all 0's except for one 1 at the x value x' (probability = .05). Let $m = 1$, so "$d(m)$" is the first element of the training set. The conjecture says that for any such $d(1)$, error must shrink. Yet if we choose $d(1)$ to be $(x', 1)$, then expected error for the Bayes-optimal algorithm for all potential test set elements $x \neq x'$ has gone from .05 to .5. So as long as $\pi(x')$ is not too large, both IID and OTS zero-one error have increased.

not specified. Yet the value of m —the size of the training set— will usually have a strong effect on the likely error. So if we don't specify those likely values of m—which is the case if we don't specify $P(m)$— we don't specify the likely effect on the error, and the learning problem is not fully defined. (When there is no noise, we do not have this difficulty. So that part of the following analysis that assumes no noise does apply to $E(C \mid m')$, even with $P(m)$ unspecified.)

Rather than entangle ourselves in considerations of what $P(m)$ should be, for simplicity here we only consider the case where we condition on m as well as m', so $P(m)$ can be unspecified. To start, let us compare $E(C \mid m', m)$ to $E(C \mid m' - 1, m - 1)$, where just as "$m - 1$" means the training set has $m - 1$ elements, "$m' - 1$" means that its $\mathcal{X}$ components contain $m' - 1$ distinct values. Unfortunately, even with m fixed, the analysis for distributions conditioned on m' seems to be intrinsically more difficult than the analysis where there is no such conditioning. Intuitively, the analysis when we only condition on m relies on associating each m-element training set d with a single $(m-1)$-element training set made up of the first $m - 1$ pairs in d. Then one shows that behavior when the algorithm sees such a d is always better than the behavior when it only sees those first $m - 1$ pairs in d. The difficulty with conditioning on m' is that there are m-element training sets, having m' distinct x values, whose first $m - 1$ values do not correspond to any $(m-1)$-element training set having $m' - 1$ distinct x values.

To illustrate this, let us try to proceed using the same arguments we did for the m-conditioned case. In analogy to Eq. 7, we can immediately write down the following:

$$E(C \mid m', m) = \sum_{d^{m-1}} P(d^{m-1} \mid m', m) \sum_{q} P(q \mid d^{m-1}_{\mathcal{X}}, m', m) E(C \mid q, d^{m-1}, m', m) \ ,$$

$$E(C \mid m' - 1, m - 1) =$$

$$\sum_{d^{m-1}} P(d^{m-1} \mid m' - 1, m - 1) \sum_{q} P(q \mid d^{m-1}_{\mathcal{X}}, m' - 1, m - 1) E(C \mid q, d^{m-1}, m' - 1, m - 1) \ . \qquad (14)$$

Unfortunately, there can be d^{m-1} in the first of our two sums that don't exist in the second. For example, say $m' < m$. Then $P(d^{m-1} \mid m' - 1, m - 1)$ tautologically equals 0 if $d^{m-1}_{\mathcal{X}}$ has m' distinct values. On the other hand, $P(d^{m-1} \mid m', m)$ need not equal 0 just because $d^{m-1}_{\mathcal{X}}$ has m' distinct values. The m'th values in $d_{\mathcal{X}}$ can be the same as one of the first $m - 1$ values.

However consider the special case where π is fixed and uniform, and there is no noise. Then by symmetry, given m', there is no additional information conveyed by specifying that

m has some particular ($\geq m'$) value. This is because under these conditions, the probability of a particular set of the m' unique elements in $d_{\mathcal{X}}$ cannot change as m is raised while m' is kept constant.[7] Accordingly, changing that value of m doesn't change expected error. For this scenario, without any loss of generality, we can always assume that all elements in the training set have distinct x values, i.e., $m' = m$.

Given such an assumption, since π is uniform, by symmetry $P(d^{m-1} \mid m'-1, m-1) = P(d^{m-1} \mid m', m)$.[8] So Eq. 14 becomes

$$E(C \mid m', m) = \sum_{d^{m-1}} P(d^{m-1} \mid m'-1, m-1) \sum_{q} P(q \mid d_{\mathcal{X}}^{m-1}, m', m) E(C \mid q, d^{m-1}, m', m) \ ,$$

$$E(C \mid m'-1, m-1) =$$

$$\sum_{d^{m-1}} P(d^{m-1} \mid m'-1, m-1) \sum_{q} P(q \mid d_{\mathcal{X}}^{m-1}, m'-1, m-1) E(C \mid q, d^{m-1}, m'-1, m-1) \ . \qquad (15)$$

Now we are in the same position in which we started our m-conditioned analysis. So our goal is to prove that $E(C \mid q, d^{m-1}, m', m) \leq E(C \mid q, d^{m-1}, m'-1, m-1)$. To do that, first use essentially the same argument that established Lemma 3.1 to establish that for the IID likelihood,

$$\sum_{d(m)} P(d(m), y_F \mid q, d^{m-1}, m', m) = P(y_F \mid q, d^{m-1}, m'-1, m-1) \ .$$

Similarly, we can prove the following analogy of Eq. 13:

$$\begin{aligned} & \sum_{y_F, y_H, d(m)} L(y_F, y_H) \ P(y_H \mid d^m, q, BO) \ P(y_F, d(m) \mid q, d^{m-1}, m, m') \\ \leq & \sum_{y_F, y_H, d(m)} L(y_F, y_H) \ P(y_H \mid d^m, q, \gamma) \ P(y_F, d(m) \mid q, d^{m-1}, m, m') \ . \end{aligned} \qquad (16)$$

Given these results, we can prove our goal inequality by using essentially the same proof which established Theorem 3.2, only expanding $E(C \mid q, d^{m-1}, m'-1, m-1)$ rather than

[7] If π were not fixed, but were instead a random variable, then changing the value of m for fixed m' would change the likely π. This in turn would change the likely values of any variable — like the error — that depends on π. Similarly, if π were fixed but not uniform, raising m for fixed m' would change the probabilities of the possible sets of distinct x values in $d_{\mathcal{X}}$.

[8] Note that this is not the case when π is non-uniform. As an example, take $n = r = 2$, and let $\pi(x = 0) = .9$. Let $m = m' = 2$, and have $d_{\mathcal{X}}^{m-1}$ equal (0). Then $P(d_{\mathcal{X}}^{m-1} \mid m'-1, m-1) = .9$. However for the $m' = m = 2$ case, both x values in the training set must be distinct, and therefore the probability that the first one has $x = 0$ is just one half.

$E(C \mid q, d^{m-1}, m-1)$. In this way establish the following theorem (where there is an implicit assumption that the random variable with values m is always taken to equal the number of distinct pairs in the training set):

Theorem 3.4 *For the IID likelihood, for uniform π and no noise, for either IID error or OTS error,*

$E(C \mid m', m, BO) \leq E(C \mid m' - 1, m - 1, BO)$.

Of course, for no noise and uniform π, we have complete freedom to fix m (see the discussion below Eq. 14). So under those conditions, Theorem 3.4 implies that $E(C \mid m' - 1, m, BO) \geq E(C \mid m', m, BO)$.

It is not currently known if Theorem 3.4 holds when noise and/or non-uniform π is allowed. However it is known that if one allows noise and/or non-uniform π, then in general this non-increasing behavior does *not* hold if one is considering the difference between $E(C \mid m'-1, m, BO)$ and $E(C \mid m', m, BO)$. This is true even for IID error. The following example illustrates this phenomenon.

Example: Let $r = 2$, $n = 3$. Assume there are two possible targets, with equal prior probability. They are given by two input-output functions, each with IID noise of 25% superimposed. (Here the "value of the noise" is the probability that at each x, rather than y being the value of the input-output function at hand, $\phi(x)$, y equals $(\phi(x) + 1) mod(2)$.) The two input-output functions have as their successive y values (0, 0, 0) and (1, 0, 0). The IID likelihood and IID $P(q \mid d)$ is assumed, as is the zero-one loss function, and $\pi(x = 0) = .999999$.

Let $m = 3$, and consider the difference between $E(C \mid m' = 1, m)$ and $E(C \mid m' = 3, m)$. Due to π, it is highly likely that the three elements of the training set contributing to $E(C \mid m' = 1, m)$ all have $x = 0$. Similarly, it is highly likely that $q = 0$. So $E(C \mid m' = 1, m)$ is given by $\sum_{d_\mathcal{Y}} E(C \mid d_\mathcal{Y}, d_\mathcal{X} = (0, 0, 0), q = 0)\ P(d_\mathcal{Y} \mid d_\mathcal{X} = (0, 0, 0))$.

To evaluate this sum, without loss of generality assume that the input-output function has the value 0 at $x = 0$. The eight possible $d_\mathcal{Y}$, (all corresponding to the same x value since $m' = 1$) are {(0, 0, 0), (0, 0, 1), (0, 1, 0), (1, 0, 0), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1)}. They have probabilities 27/64, 9/64, 9/64, 9/64, 3/64, 3/64, 3/64, and 1/64, respectively, given our chosen target. For the first four of these, since there are more 0's than 1's in $d_\mathcal{Y}$, the Bayes-optimal guess will be 0. For such a guess, the expected error is 1/4 (that's the probability that the test set element at $x = 0$ has the value $y = 1$). For the last four, the Bayes-optimal guess will be 1, resulting in expected error 3/4. Performing the sum, the expected error equals $84/256 = 21/64$.

Now consider the case where $m' = 3$, i.e., where all three elements of the training set have different x's. Again the test set element will with high probability be at $x = 0$. Without loss

of generality, assume the input-output function has the value 0 there. Then the probability is 3/4 that at $x = 0$ the training set value is $y = 0$, and the probability that $y = 1$ there is 1/4. In the former case, the Bayes-optimal guess at $x = 0$ is 0, resulting in expected error 1/4, and in the latter case, the guess is 1, resulting in expected error 3/4. Performing the sum, we get $6/16 = 24/64$, which exceeds 21/64. **QED**

3.3 m'-dependent Behavior when the Sampling Distribution is Uniform

Before considering in detail OTS learning curves for uniform π, it is necessary to establish some elementary identities. That is done in this subsection.

Consider the trivial learning algorithm that always makes the same guess g, independent of the data. The following lemma formalizes the (intuitively reasonable) statement that for this learning algorithm, $E(C|m')$ is independent of m'. The lemma holds for any $P(f)$, any noise process (i.e., any $P(d_Y|f, d_X)$), and any loss function.

Lemma 3.5 *For uniform $\pi(x)$ and the IID likelihood, if $P(h \mid d) = \delta(h - g(x))$ for some function $g(x)$, then $E(C \mid m')$ is independent of m'.*

Proof: First note that

$$\begin{aligned} E(C \mid m') &= \sum_{f,h,d} E(C \mid f, h, d) P(f, h, d) \\ &= \sum_{f,h,d} \left\{ P(d_{\mathcal{Y}} \mid d_{\mathcal{X}}, f) P(d_{\mathcal{X}} \mid f) P(f) P(h \mid d) \left[\frac{\sum_{x \notin d_{\mathcal{X}}} \pi(x) L(f(x), g(x))}{\sum_{x \notin d_{\mathcal{X}}} \pi(x)} \right] \right\} \end{aligned}$$

In these sums it is implicitly assumed that we are only considering those d having the appropriate value of m', and that $P(f, h, d)$ is normalized accordingly. (Similar implicit assumptions are made below.) Write the outer sum in our expression for $E(C \mid m')$ as $\sum_{f,h,d_{\mathcal{X}},d_{\mathcal{Y}}}$ and eliminate the $P(d_{\mathcal{Y}} \mid d_{\mathcal{X}}, f)$ by performing the sum over $d_{\mathcal{Y}}$. Then use the assumption that $P(d_{\mathcal{X}} \mid f) = P(d_{\mathcal{X}}) = \prod_i \pi(d_{\mathcal{X}}(i))$. (Recall the definition of the IID likelihood.) Next use $P(h \mid d) = \delta(h - g)$ and take $\pi(x) = 1/n$:

$$E(C \mid m') = \sum_{f,d_{\mathcal{X}},x \notin d_{\mathcal{X}}} L(f(x), g(x)) P(d_{\mathcal{X}}) P(f) \; / \; (n - m') \tag{17}$$

$$= \sum_{f,x} L(f(x), g(x)) P(f) \left[\sum_{d_{\mathcal{X}}: x \notin d_{\mathcal{X}}} \frac{P(d_{\mathcal{X}})}{(n - m')} \right] \tag{18}$$

To evaluate the term in the [..], we must change variables. $d_{\mathcal{X}}$ is an ordered set of x values, and can contain duplicates in general. Let $d'_{\mathcal{X}}$ be the associated unordered set of x values with all duplicates removed. The mapping from the space of $d_{\mathcal{X}}$'s to $d'_{\mathcal{X}}$'s is single-valued. Moreover, the set of all $d_{\mathcal{X}}$ s.t. $(x \notin d_{\mathcal{X}})$ maps to the set of all $d'_{\mathcal{X}}$ s.t. $(x \notin d'_{\mathcal{X}})$. Therefore we can rewrite the [..] as

$$\sum_{d'_{\mathcal{X}}: x \notin d'_{\mathcal{X}}} P(d'_{\mathcal{X}}) \ / \ (n - m')$$

(Recall that m' is implicitly fixed.)

Now since $\pi(x)$ is uniform, by symmetry $P(d'_{\mathcal{X}})$ is the same for all terms in this sum. Therefore by normalization, for any $d'_{\mathcal{X}}$ in the sum $P(d'_{\mathcal{X}})$ equals $\binom{n}{m'}^{-1}$. The number of terms in the sum is $\binom{n-1}{m'}$, due to the $x \notin d'_{\mathcal{X}}$ condition. Therefore the sum equals $1/n$, and

$$E(C \mid m') = \frac{1}{n} \sum_{f,x} L(f(x), g(x)) \ P(f)$$

independent of m'. **QED.**

It may seem obvious that the expected cost of the generalizer that always guesses the same function, no matter what the data, doesn't vary with the size of that data. However this is *not* true in general for non-uniform $\pi(x)$ and OTS error. An example is that scenario presented at the end of the previous subsection in which the OTS error of the Bayes optimal generalizer increases with m'.

The fact that $\pi(x)$ being uniform is important both for OTS error to be non-increasing and for Lemma 3.5 is no coincidence; the two results are closely related, as the following lemma shows:

Lemma 3.6 *For uniform $\pi(x)$, no noise, the IID likelihood and $0 < k < n$, $E(C \mid m' = k, BO) = E(C \mid m' = 0, BO)$ iff there exists a data-independent $g(x)$ such that $E(C \mid m' = k, BO) = E(C \mid m' = k, g)$, where the g-conditioning means a generalizer $P(h \mid d) = \delta(h - g)$.*

Proof: Let "$(BO(m' = z))$" mean the generalizer that sees only the first z distinct elements in d and is Bayes-optimal given those elements. Then the phrase "$E(C \mid m' = a, BO)$" in the statment of the Lemma means $E(C \mid m' = a, BO(m' = a))$.

Now if $E(C \mid m' = k, BO(m' = k)) = E(C \mid m' = 0, BO(m' = 0))$, then since $BO(m' = 0)$ is independent of any data, by Lemma (3.5)

$$E(C \mid m' = k, BO(m' = k)) = E(c \mid m' = k, BO(m' = 0)).$$

Set the guess made by the generalizer $BO(m' = 0)$ to be $g(x)$. Then our equality implies that there exists a $g(x)$ s.t. $E(C \mid m' = k, BO) = E(C \mid m' = k, g)$.

Going in the other direction, if there exists a $g(x)$ such that

$$E(C \mid m' = k, BO) = E(C \mid m' = k, g),$$

then by Theorem 3.4 $E(C \mid m' = k, g) \leq E(C \mid m' = 0, BO)$. But by Lemma (3.5) and Bayes-optimality,

$$E(C \mid m' = k, g) = E(C \mid m' = 0, g) \geq E(C \mid m' = 0, BO).$$

Therefore $E(C \mid m' = k, g) = E(C \mid m' = 0, BO)$. Therefore (by our hypothesis)

$$E(C \mid m' = k, BO) = E(C \mid m' = 0, BO). \qquad \textbf{QED}$$

Any generalizer that minimizes $E(C \mid m)$ necessarily makes the same guesses as the Bayes-optimal generalizer for all d such that $P(d) \neq 0$ [22]. (Recall that if more than one h minimize $E(C \mid d)$ for some d, then saying a generalizer is Bayes-optimal simply means that it guesses one of those optimal h in response to d.) Accordingly, Lemma 3.6 tells us that $E(C \mid m' = k, BO) = E(C \mid m' = 0, BO)$ if and only if one Bayes-optimal generalizer guesses some function $g(x)$ in response to all allowed training sets.

3.4 Constant OTS Error and Ideals

An interesting problem is to determine the conditions under which the OTS error is constant for all $m' < n$ (the m' for which OTS error is defined). For the analysis of this issue presented in this subsection, we assume that $\mathcal{Y}$ is binary, there is no noise, and $\pi(x)$ is uniform. Because there is no noise, all elements in the training set with the same $\mathcal{X}$ value have the same $\mathcal{Y}$ value. Accordingly, knowing m in addition to m' conveys no extra information, so if m' is specified in the conditioning event there is no reason to specify m as well. Moreover, in this scenario, all targets f are vectors living at the vertices of the hypercube $2^{[n]}$. In addition, the only d such that $P(d) = 0$ are those that lie on an f for which $P(f) = 0$. For $u, v \in 2^{[n]}$, let $u \oplus v$ denotes the pointwise exclusive OR (or the difference vector) between u and v. Let $u \geq v$ mean that $u_i \geq v_i$ for every i.

Definition: Let P be a probability distribution on $2^{[n]}$. P is a *relative ideal* iff there exists a $g \in 2^{[n]}$ (the *center* of P) such that for all $u, v \in 2^{[n]}$ with $u \oplus g \geq v \oplus g$, $P(u) \leq P(v)$.

A simple example of a relative ideal is a P that is constant for all vectors within some Hamming distance R of g, and zero otherwise. Another example is the $P(f)$ of the example at the end of Section 2.3.

The following theorem relates relative ideals to constancy of the learning curve.

Theorem 3.7 *Under the assumptions stipulated at the beginning of this subsection, $E(C \mid m', BO)$ is constant for all $m' \leq n-1$ iff $P(f)$ is a relative ideal.*

Proof: Suppose that $E(C \mid m', BO)$ is constant for all m' up to $m' = n-1$. Then by the discussion following Lemma 3.6, one Bayes optimal generalizer (there may be more than one) guesses the same function $g \in 2^{[n]}$ in response to any allowed training set having $m' < n$. Suppose that t is the function defined by a training set d having $m' = n-1$. (The domain of t is $d_{\mathcal{X}}$.) Let i be the only input not seen. The only targets consistent with d are the two extensions of t defined by $t_i^{(0)} \equiv g_i$ and $t_i^{(1)} \equiv g_i \oplus 1(i)$. Since g is a Bayes optimal hypothesis, $P_F(t^{(1)}) \leq P_F(t^{(0)})$. This inequality holds no matter what t is (so long as it lies on an f with non-zero prior probability), and no matter what input remains to be seen.

Now suppose that $u \oplus g \geq v \oplus g$, and that $P_F(u) \neq 0$. (If $P_F(u) = 0$, then it trivially follows that $P_F(u) \leq P_F(v)$.) Since $P_F(u) \neq 0$, any training set lying on u is allowed. Furthermore, since $u \oplus g \geq v \oplus g$, there is a "chain" $v^{(0)} = u, v^{(1)}, \ldots, v^{(k)} = v$ such that $v^{(i)} \oplus g \geq v^{(i+1)} \oplus g$ and $v^{(i)}$ and $v^{(i+1)}$ differ in only one position. This implies that $P_F(v^{(i)}) \leq P_F(v^{(i+1)})$ for each i (use the argument of the preceding paragraph with t defined to be the $n-1$ points on which $v^{(i)}$ and $v^{(i+1)}$ agree). Hence, $P_F(u) \leq P_F(v)$.

To prove the other direction, suppose that P is a relative ideal with center g and that we are given a d with $m' = n-1$. As in the previous paragraph, let t be the function defined by d, let i be the sole element of $\mathcal{X}$ not in $d_{\mathcal{X}}$, and define the two extensions of t by $t_i^{(0)} \equiv g_i$ and $t_i^{(1)} \equiv g_i \oplus 1(i)$. The inequality of the definition of relative ideals implies that g is a Bayes optimal guess for input i (choose $v = t^{(0)}$ and $u = t^{(1)}$). Since this is true for any d, it is also true when we average over d's. The inequality also implies that g is the Bayes optimal guess when no training examples have been seen (choose $v = g$). So the Bayes-optimal generalizer is identical to the generalizer of Lemma 3.5 for these two m' values, for points outside of the training set. Accordingly, by Lemma 3.5, $E(C \mid m', BO)$ is the same for those two m' values. Therefore by Theorem 3.4, it is independent of m'. **QED**

Next consider $E(C \mid f^\star, m', BO)$, where $f^\star$ is some target with non-zero prior probability. Even for a uniform $\pi(x)$ this expected cost can increase with increasing m' [22]. However the following corollary of Theorem 3.7 allows us to rule out such behavior in many cases.

Corollary 3.8 *Under the assumptions stipulated at the beginning of this subsection, for any target $f^\star$ whose prior probability does not equal zero, $E(C \mid f^\star, m', BO)$ can vary with m' only if $P(f)$ is not a relative ideal.*

Proof: By Theorem 3.7, if $P(f)$ is a relative ideal, $E(C \mid m', BO)$ is a constant function of m'. By Lemma 3.6, that would imply that a Bayes-optimal generalizer (there may be more than one) always guesses the same function $g(x)$ in response to any allowed training set.

This in turn would mean that the value of $E(C \mid f^\star, m', BO)$ for the actual (relative ideal) prior at hand is equal to the value of $E(C \mid m', g)$ under the prior $P(f) = \delta(f - f^\star)$. By Lemma 3.5 however this latter expression cannot vary with m'. **QED**

We can use Corollary 3.8 to derive the following result:

Corollary 3.9 *Under the assumptions stipulated at the beginning of this subsection, $E(C \mid m', BO)$ is a non-constant function of m' iff there is variability over m' in the values of $E(C \mid f, m', BO)$ for some f with a non-zero prior probability.*

Proof: $E(C \mid m', BO) = \sum_f E(C \mid f, m', BO)\ P(f)$. Therefore if $E(C \mid m', BO)$ varies with m', the same must be true for $E(C \mid f, m', BO)$ for some f. Conversely, if $E(C \mid m', BO)$ is independent of m', then by Theorem 3.7, $P(f)$ is a relative ideal. So by Corollary 3.8, $E(C \mid f, m', BO)$ is independent of m' for all f with non-zero prior. **QED**

So by Theorem 3.4, $E(C \mid f^\star, m', BO)$ can increase with m' only if $E(C \mid m', BO)$ decreases with m'.

3.5 Results for Gibbs Generalizer Learning Curves

Eq. 7 and Lemma 3.1 control the learning curves for Gibbs generalizers just as they do for Bayes generalizers. This is established in the following theorem.

Theorem 3.10 *If $L(y, y')$ is a negative semi-definite matrix over the subspace perpendicular to $\vec{1}$,*[9] *then $E(C \mid q, d^{m-1}, m, G) \leq E(C \mid q, d^{m-1}, m-1, G)$.*

Proof: Define $v(y_F, q, d) = P(y_F \mid q, d, m)$. Now

$$E(C \mid q, d^{m-1}, m, G) = \sum_{y, y', d(m)} L(y, y')\, v(y, q, d^m)\, v(y', q, d^m)\, P(d(m) \mid q, d^{m-1}, m) \qquad (19)$$

for a correct Gibbs generalizer (recall Eq. 12). This can be rewritten as

$$E(C \mid q, d^{m-1}, m, G) =$$

$$\sum_{y, y', d(m), d'(m)} L(y, y')\, v(y, q, d^m)\, v(y', q, d^m)\, P(d(m) \mid q, d^{m-1}, m)\, P(d'(m) \mid q, d^{m-1}, m) \qquad (20)$$

where $d'(m)$ is a potential m-th element of d, not necessarily the same as $d(m)$.

[9] I.e. $\sum_{y,y'} L(y, y') a_y a_{y'} \leq 0$ whenever $\sum_y a_y = 0$.

On the other hand,

$$E(C \mid q, d^{m-1}, m-1, G) = \sum_{y_F, y_H} L(y_F, y_H)\, P(y_F \mid q, d^{m-1}, m-1)\, P(y_H \mid q, d^{m-1}, m-1, G)\ .$$

Now use Lemma 3.1 to get

$$E(C \mid q, d^{m-1}, m-1, G) =$$

$$\sum_{y, y', d(m), d'(m)} L(y, y')\, v(y, q, d^m)\, v(y', q, d'^m)\, P(d(m) \mid q, d^{m-1}, m)\, P(d'(m) \mid q, d^{m-1}, m)\ , \quad (21)$$

where $d'^m \equiv \{d^{m-1}, d'(m)\}$.

The only difference between the expression for $E(C \mid q, d^{m-1}, m-1, G)$ and the one for $E(C \mid q, d^{m-1}, m, G)$ is whether the summand contains $v(y', q, d^m)$ or $v(y', q, d'^m)$. It is this difference that establishes the theorem. To see this write the difference as

$$\sum_{y, y', d(m), d'(m)} L(y, y')\, v(y, q, d^m)\, P(d(m) \mid q, d^{m-1}, m)\, P(d'(m) \mid q, d^{m-1}, m)\, [v(y', q, d'^m) - v(y', q, d^m)]\ .$$

Now rewrite this expression by interchanging $d(m)$ and $d'(m)$. Add our two expressions for $E(C \mid q, d^{m-1}, m-1, G) - E(C \mid q, d^{m-1}, m, G)$ and divide by 2. The result is that the difference $E(C \mid q, d^{m-1}, m-1, G) - E(C \mid q, d^{m-1}, m, G)$ is proportional to

$$- \sum_{d(m), d'(m)} P(d(m) \mid q, d^{m-1}, m) P(d'(m) \mid q, d^{m-1}, m) \sum_{y, y'} L\,(y, y')\, [v(y, d^m, q) - v(y, d'^m, q)]\, [v(y', d^m, q) - v(y', d'^m, q)]\ . \quad (22)$$

View the [...] terms as the same vector over $\mathcal{Y}^n$, indexed either by y or y' Since it is the difference of two probability distributions, that vector is perpendicular to the "one" vector, $\vec{1} \equiv (1, 1, ...)$. However by hypothesis L is negative semidefinite on this subspace. Therefore the sum over y and y' is negative semidefinite for any $d(m)$ and $d'(m)$. Accordingly the full sum is positive and $E(C \mid q, d^{m-1}, m-1, G) \geq E(C \mid q, d^{m-1}, m, G)$. **QED**

Now apply Eq. 7 in the usual way. The result is that for loss functions of the type specified in Theorem 3.10, for either IID error or OTS error with uniform π, expected error is non-increasing with training set size, along average trajectories, etc.

Both the zero-one and the quadratic loss functions have the property specified in Theorem 3.10. Intuitively, those loss functions for which Theorem 3.10 does not apply and the Gibbs generalizer's learning curve increases are those for which the loss shrinks as h and f get *further* apart. Interestingly, the learning curve is non-increasing for the Bayes-optimal generalizer even for such a "backwards" loss function.

3.6 Relating IID and OTS Error

Since IID error and OTS error are closely related, results concerning the one have implications for the other. To see this, with a bit of abuse of notation and with the dependencies of $\bar{C}$ made explicit, use Eq. 3 to write

$$\bar{C}_{IID}(f,h,d) = \pi(\mathcal{X} - d_{\mathcal{X}})\bar{C}_{OTS}(f,h,d) + \pi(d_{\mathcal{X}})\bar{C}_{TS}(f,h,d) \tag{23}$$

where the subscript TS means "on the training set" and $\pi(A)$ means $\sum_{x \in A} \pi(x)$.

So for example consider the case where there is no noise (so $\bar{C}_{TS}(f,h,d) = 0$ for the Bayes-optimal generalizer). Write

$$\begin{aligned}
\bar{C}_{OTS} \equiv E(C_{OTS} \mid m) &= \sum_{f,h,d} E(C_{OTS} \mid f,h,d)\ P(f,h,d \mid m) \\
&= \sum_{f,h,d} \bar{C}_{OTS}(f,h,d)\ P(f,h,d \mid m) \\
&= \sum_{f,h,d} \frac{\bar{C}_{IID}(f,h,d)\ P(f,h,d \mid m)}{\pi(\mathcal{X} - d_{\mathcal{X}})} \\
&= \sum_{d} \bar{C}_{IID}(d)\ P(d_{\mathcal{Y}} \mid d_{\mathcal{X}}, m)\ \frac{\pi(d_{\mathcal{X}})}{\pi(\mathcal{X} - d_{\mathcal{X}})}\ ,
\end{aligned}$$

where Eq. 3 is used to establish the last step and we define $\bar{C}_{IID}(d) = \sum_{f,h} \bar{C}_{IID}(f,h,d) P(f,h \mid d)$.

Combined with our result that $\bar{C}_{IID} = \sum_d \bar{C}_{IID}(d)\ P(d_{\mathcal{Y}} \mid d_{\mathcal{X}}, m)\ \pi(d_{\mathcal{X}})$ is non-increasing as a function of m for all $\pi(x)$ for the Bayes-optimal generalizer (Theorem 3.3), this controls how much $\bar{C}_{OTS}$ can increase for the Bayes-optimal generalizer for non-uniform $\pi(x)$.

As another example, in the specific cases where we can limit the OTS error, we can use Eq. 23 to limit the IID error. In particular, for uniform $\pi(x)$ and no noise, $E(C_{TS}) = 0$ for both the Bayes Optimal and the Gibbs generalizers (with correct prior), and we have

$$E(\bar{C}_{IID} \mid m') = \frac{n - m'}{n}\ E(\bar{C}_{OTS} \mid m')\ . \tag{24}$$

Since $E(\bar{C}_{OTS} \mid m') = E(C_{OTS} \mid m')$ is non-increasing for this case for both generalizers (assuming the relevant conditions outlined above are met), we have a bound on how slowly the value $E(\bar{C}_{IID} \mid m')$ can decrease with m'. The worst-case is where $E(C_{OTS} \mid m')$ is constant (which means we have a *relative ideal* - see above). In that case, we have

$$E(\bar{C}_{IID} \mid m) = [1 - \frac{E(m' \mid m)}{n}]\ E(\bar{C}_{OTS} \mid m)\ .$$

For uniform π, $E(m' \mid m) = n(1 - (1 - \frac{1}{n})^m)$. (Use the fact that the probability that any particular element of $\mathcal{X}$ never occurs in the m elements is $(1 - \frac{1}{n})^m$, and the fact that the expectation value of a sum is a sum of expectation values.) For $n \gg m$ this gives

$$E(\bar{C}_{IID} \mid m) \approx [1 - (\frac{m}{n})] \; E(\bar{C}_{OTS} \mid m) \; .$$

Consider the case where $r = 2$. Together with no noise and zero-one loss, this is the scenario considered in [7]. (Though note that for $r = 2$, results pertaining to zero-one loss also hold for quadratic loss.) To illustrate the difference between the bounds derived here and those in [7], let $P(f)$ be uniform over all f with one or fewer 1's and let π be uniform. For this case it is straight-forward to show that we have a relative ideal, so $E(\bar{C}_{OTS} \mid m')$ is independent of m', and therefore can be evaluated as $1/(n+1)$, by taking $m' = 0$. So for $n \gg m$ we get

$$E(\bar{C}_{IID} \mid m) \approx [\frac{1}{n+1} - \frac{m}{n(n+1)}] \; . \tag{25}$$

For this case, the number of possible targets is given by $|F| = n + 1$. So the bound on $E(\bar{C}_{IID} \mid m)$ given in [7], $\frac{|F|}{m}$, is $(n+1)/m$. This is a very poor bound on the value given in Eq. 25. (Note also that the result in [7] doesn't actually show that $E(\bar{C}_{IID} \mid m)$ is shrinking, merely that an upper bound on it is.) Part of the reason for this poorness of the bound in [7] is that it holds for all $P(f)$ and $\pi(x)$.

4 The Off-Training-Set-Error for the Circuit Case

In this section we illustrate our main OTS results by solving exactly a simple example where the target distribution $P_F(f)$ is a *circuit* in the space of boolean functions, there is no noise, the sampling distribution is uniform, and we have the zero-one loss function. (See the comments on this scenario at the beginning of the subsection on relative ideals.)

Since we have an input space $\mathcal{X}$ of size $|\mathcal{X}| = n$, the set of all boolean functions over that space is the n dimensional hypercube $\{0,1\}^{\mathcal{X}}$. We consider a target distribution $P_F(f)$ which is uniform over a set of boolean functions called the target class U and zero outside that class (see the example in Section 2.3). The *circuit* target class[10], denoted by $U = Circ(\vec{0}, R)$, is the set of boolean functions f for which $f(x) = 1$ for exactly R elements of $\mathcal{X}$. It can be represented by the set of binary vectors v, of size n and weight $|v| = R$. All the calculations we are doing here are the same for the class $Circ(\vec{w}, R)$, the circuit of radius R around an arbitrary center $\vec{w}$, since the two classes are obtained from each other by the "partial-flipping" automorphism operation on the hypercube.

[10] Also known as a *Hamming sphere*.

The definition of Bayes Optimal (BO) "guessing" in this context is defined as follows:

- Given a target class U which is a set of functions in $\{0,1\}^{\mathcal{X}}$.
- Given a training set d which is a set of pairs $(x \in \mathcal{X}, f(x))$ for a predetermined $f \in U$.
- Given a question $q \in \mathcal{X}$ (but $q \notin d_{\mathcal{X}}$), the BO guess for $f(x)$ is given by counting the number of elements in $U(d, f(q) = 1)$ (the set of functions $f \in U$ that are consistent with the training examples d and for which $f(q) = 1$), and counting the number of elements in $U(d, f(q) = 0)$ (the set of consistent functions in U for which $f(q) = 0$). Its output is

$$BO(U, d, q) = \begin{cases} 1 & \text{if } |U(d, f(q) = 1)| > |U(d, f(q) = 0)| \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

We shall calculate the expected OTS error of the BO and Gibbs algorithm for the specific choice of the circuit target class U. We will assume uniform distributions over U and $\mathcal{X}$ (i.e., we assume that $P(f)$ is constant over its support, and that $\pi(x) = 1/n$ for all x).

4.1 The Case $m = 0$

We start with the simplest case where there is no training data, i.e. $m = 0$. We then calculate the expected error $E(C \mid m', m, U = Circ)$ for any m' and m.

For every $q \in \mathcal{X}$, it is easy to verify that

$$|U(f(q) = 1)| = \binom{n-1}{R-1} \quad \text{, the number of functions in } Circ(\vec{0}, R) \text{ for which } f(q) = 1.$$

$$|U(f(q) = 0)| = \binom{n-1}{R} \quad \text{, the number of functions in } Circ(\vec{0}, R) \text{ for which } f(q) = 0.$$

Without loss of generality we assume that $R \leq n/2$. (If this isn't the case we can make it so by just interchanging 0 and 1 in $\mathcal{Y}$.) This implies $\binom{n-1}{R-1} \leq \binom{n-1}{R}$, and therefore $BO(m = 0, q) = 0$ for any q. This guess will be correct for $n - R$ of the inputs, which means that

$$E(C \mid f, m = 0) = R/n \quad , \quad \forall f \in Circ(\vec{0}, R) \tag{27}$$

which in turn implies that $E(C \mid m = 0, U = Circ) = R/n$.

4.2 The Case $0 < m < n$

With the target function $f \in U$ fixed, we can characterize the training set d by the number of $x \in d_{\mathcal{X}}$ for which $f(x) = 1$, denoted by m_1, and by the number of $x \in d_{\mathcal{X}}$ for which $f(x) = 0$, denoted by $m_0 \equiv m' - m_1$. In what follows we shall use the two variables m' and m_1 to define the training set (i.e., m is assumed fixed, and will therefore usually only be implicit). When $U = Circ(\vec{0}, R)$, $m_1 \leq R$ and $m_0 \leq n - R$. In addition, the number of functions in $Circ(\vec{0}, R)$ consistent with d is $\binom{n-m'}{R-m_1}$ ($R - m_1$ is the number of 1's to be fixed in $\mathcal{X} - d_{\mathcal{X}}$).

Choosing any $q \notin d_{\mathcal{X}}$:

$$|U(m', m_1, f(q) = 1)| = \binom{n - m' - 1}{R - m_1 - 1} \tag{28}$$

$$|U(m', m_1, f(q) = 0)| = \binom{n - m' - 1}{R - m_1} \tag{29}$$

As long as $R - m_1 \leq \frac{1}{2}(n - m')$, $|U(m', m_1, f(q) = 1)| \leq |U(m', m_1, f(q) = 0)|$. $BO(q) = 0$ for any $q \notin d_{\mathcal{X}}$ iff we are in that situation. For fixed m', this condition for $BO(q) = 0$ amounts to

$$m_1 \geq R - \left\lceil \frac{n - m' - 1}{2} \right\rceil \tag{30}$$

This always holds when $m' \leq n - 2R - 1$ (or $m' \leq n - 2R$ if m' and n are of the same parity). Above that size, there are choices of d which yield $BO(q) = 1$.

Now consider a case in which f and d are such that $BO(q) = 0$ (i.e., a case where m' and m_1 are fixed with values giving $BO(q) = 0$). In that case the guess is wrong for exactly $R - m_1$ test questions $q \notin d_{\mathcal{X}}$. Therefore in that case the expected OTS error is given by

$$E(C \mid f, m', m_1; BO = 0) = \frac{R - m_1}{n - m'} \ . \tag{31}$$

Similarly,

$$E(C \mid f, m', m_1; BO = 1) = \frac{n - R - m' + m_1}{n - m'} = 1 - \frac{R - m_1}{n - m'} \ . \tag{32}$$

Consequently, we can use Eq.30 to write

$$E(C \mid f, m', m_1) = \begin{cases} \frac{R - m_1}{n - m'} & \text{if } m_1 \geq R - \left\lceil \frac{n - m' - 1}{2} \right\rceil . \\ 1 - \frac{R - m_1}{n - m'} & \text{otherwise} \end{cases} \tag{33}$$

In order to use this to calculate $E(C \mid f, m')$ we have to find $P(m_1 \mid f, m')$. Due to the symmetry of the problem, this probability is independent of the choice of $f \in Circ(\vec{0}, R)$. Now let $\chi(m, m')$ be the number of strings of length m with exactly m' distinct letters. Then

there are $\chi(m,m')\binom{R}{m_1}\binom{n-R}{m'-m_1}$ ways to form an (ordered) training set of m elements, m' of which are distinct, and for m_1 of which $f=1$. The number of ways to form a training set of m elements, m' of which are distinct, is

$$\chi(m,m')\binom{n}{m'} = \chi(m,m')\sum_{m_1=0}^{m'}\binom{R}{m_1}\binom{n-R}{m'-m_1} \tag{34}$$

This sum is valid even when $m' > R$. Since we are assuming a uniform distribution over $\mathcal{X}$, the desired probability is

$$P(m_1 \mid f, m') = \frac{\binom{R}{m_1}\binom{n-R}{m'-m_1}}{\binom{n}{m'}} \tag{35}$$

Since both $E(C \mid f, m', m_1)$ and $P(m_1 \mid f, m')$ do not depend on f, no explicit averaging over f is needed and the expected OTS error is given by

$$E(C \mid U = Circ, m', m) = \sum_{m_1=0}^{m'} E(C \mid f, m', m_1) P(m_1 \mid f, m') \ , \tag{36}$$

where $E(C \mid f, m', m_1)$ and $P(m_1 \mid f, m')$ are given above.

Note that in principle, the upper limit of the sum should be $\mathrm{Min}(m, R)$, but using m is still valid since the binomial coefficients in $P(m_1 \mid f, m')$ are zero when $m_1 > R$. Similar considerations apply for identities (34) and (38).

4.3 The Case $BO = 0$

As mentioned above, when $m' \leq m^\star \equiv n - 2R - 1$ (or $m' \leq n - 2R$, if m' and n are of the same parity), the guess of the BO algorithm is always 0. This allows some simplifications; for this region of m' values,

$$E(C \mid U = Circ, m', m) = \binom{n}{m'}^{-1} \sum_{m_1=0}^{m'} \binom{R}{m_1}\binom{n-R}{m'-m_1}\frac{R-m_1}{n-m'} \tag{37}$$

(see Eq.33). Note that the right hand side of this equation does not depend on m.

We can evaluate this sum explicitly by using the following identity[11]

$$t\binom{s+t-1}{p-1} = \sum_{i=0}^{p} i\binom{t}{i}\binom{s}{p-i} \ . \tag{38}$$

[11] It can be proved, for example, by considering the following equality, which must hold for all x:

$$\left(\sum_i i\binom{t}{i}x^{i-1}\right)\left(\sum_j \binom{s}{j}x^j\right) = t(1+x)^{t-1}(1+x)^s = t\sum_k \binom{t+s-1}{k}x^k \ .$$

Doing this gives

$$E(C \mid U = Circ, m' \leq m^\star, m) = \frac{R}{n-m'} - \frac{1}{(n-m')\binom{n}{m'}} R \binom{n-1}{m'-1} \quad (39)$$

$$= \frac{R}{n-m'}\left(1 - \frac{m'}{n}\right) \quad (40)$$

$$= \frac{R}{n} = E(C \mid U = Circ, m' = 0, m) \; . \quad (41)$$

So we see that the average off-training-set error remains constant for m' ranging from 0 up to $m^\star$.

We can also show that for $m' > m^\star$, value $E(C \mid U = Circ, m', m)$ decreases. To see this, consider the case where $BO(q) = 1$ (see the condition in Eq.30). In this case $R - m_1 > \left\lceil \frac{n-m'-1}{2} \right\rceil$, and if we take $n - m'$ even for example, we find that

$$\frac{R-m_1}{n-m'} > \frac{1}{2} > 1 - \frac{R-m_1}{n-m'} \; . \quad (42)$$

This means that in the sum of Eq.37, some of the terms are replaced by smaller terms. Accordingly $E(C \mid U = Circ, m' > m^\star, m) < E(C \mid U = Circ, m' \leq m^\star, m)$. The full curves of $E(C \mid U = Circ, m', m, BO)$ vs. m', for $n = 128$ and $R = 32, 60$, are shown in Figure 1, along with the Gibbs generalizer learning curves for those scenarios.

4.4 The Large n Limit

To evaluate the results obtained so far in the large n limit with m'/n and R/n held constant (the *thermodynamic limit*), it is only necessary to replace the sum over m_1, which is the average over the realizations of the training set, by an integral. In that situation $P(m_1 \mid f, m)$ (Eq.35), is well approximated by a Gaussian distribution centered around $m_1 = m'R/n$. In the limit $n \to \infty$, since we are keeping the ratio $R/n \equiv r$ fixed, this Gaussian becomes a delta function. Thus in the thermodynamic limit we obtain the following:

$$E(C \mid U = Circ, m', m, BO) \approx \frac{R - \frac{m'R}{n}}{n-m'} = r \quad (43)$$

for $r \leq 1/2$ (cf. Eq. 33; $r \leq 1/2$ ensures we're in the upper condition).

This means that in the large n limit the OTS error remains almost constant as $m'/n \to 1$. This is not too surprising. After all, in the $n \to \infty$ limit, the circuit with radius R on the binary hypercube contains almost all the volume of the sphere with that radius. As a result the error for the circuit prior behaves like the error for the prior that's uniform over all targets inside the sphere. In turn, the error for the sphere prior is constant (since the sphere is a relative ideal).

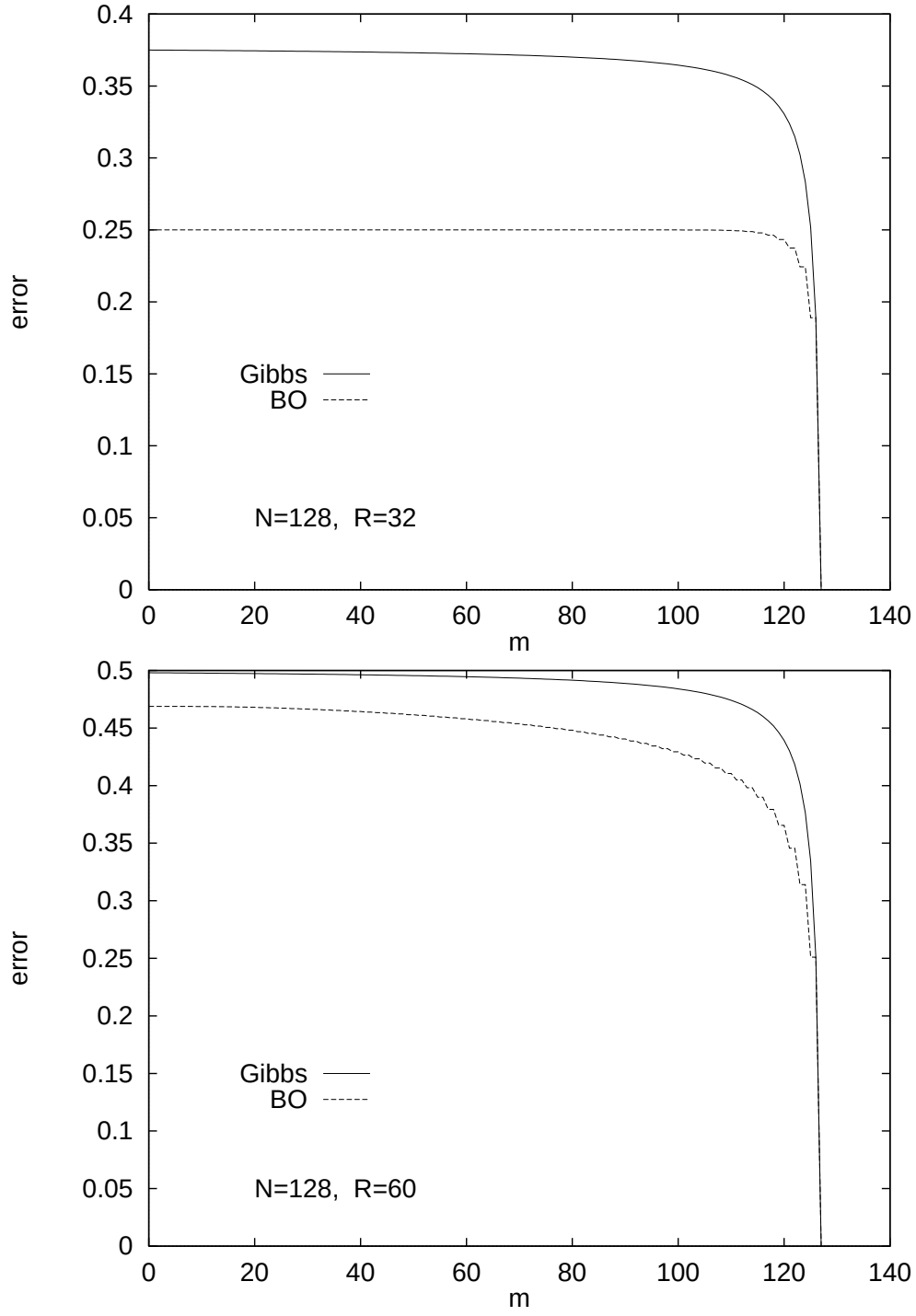

Figure 1: The expected OTS error of the Gibbs and Bayes Optimal algorithms for the circuit prior as a function of m'. $n = 128, R = 32$ (top) and $n = 128, R = 60$ (bottom).

4.5 The Gibbs Algorithm Calculation

Consider now $E(C \mid U = Circ, m', m, G)$, the expected OTS for the Gibbs algorithm with the circuit prior. As before, the training set is characterized by m' and by m_1, the number of examples with $f(x) = 1$. Now using the Gibbs algorithm to choose h from the R circuit hypothesis space is equivalent to choosing $R - m_1$ additional x values (beyond any in $d_{\mathcal{X}}$) for which $h(x) = 1$. Given any target function $f \in Circ(\vec{0}, R)$ and any training set d sampled from f, randomly choosing these $R - m_1$ x values from the $n - m$ undetermined x lying outside of $d_{\mathcal{X}}$ will result in l_1 of those values corresponding to x's such that $f(x) = 1$. The remaining $l_0 \equiv R - m_1 - l_1$ of the $R - m_1$ x values will have $f(x) = 1$, and therefore will be in error. As a result, the OTS error is just $2l_0$ (for any x where $h(x)$ incorrectly equals 1, there is also one x for which $h(x)$ incorrectly equals 0). So the average error over q is $\frac{2l_0}{n-m}$.

This q-average now must be averaged over the possible choices of h (characterized by l_0). Assuming a uniform sampling distribution over $\mathcal{X}$, we only have to consider the combinatorics of l_0. Since l_0 is the number of x such that $h(x) = 1$ but $f(x) = 0$ that are uncovered by the training set, it is limited to the region $0 \leq l_0 \leq l'_m \equiv Min\,(R - m_1, n - m' - (R - m_1))$.

So the average over h is given by

$$E_G(m', m_1) \equiv \frac{1}{n - m'} \binom{n - m'}{R - m_1}^{-1} \sum_{l_0=0}^{l_{m'}} 2l_0 \binom{n - R - m_0}{l_0} \binom{R - m_1}{l_1} \tag{44}$$

$$= \frac{2}{n - m'} \binom{n - m'}{R - m_1}^{-1} \sum_{l_0=0}^{l_{m'}} l_0 \binom{n - R - m' + m_1}{l_0} \binom{R - m_1}{R - m_1 - l_0} . \tag{45}$$

Using the identity (38) again, this can be evaluated to give

$$E_G(m', m_1) = \frac{2}{n - m'} \binom{n - m'}{R - m_1}^{-1} (n - R - m' + m_1) \binom{n - m' - 1}{R - m_1 - 1} = 2 \frac{R - m_1}{n - m'} \left[1 - \frac{R - m_1}{n - m'}\right] . \tag{46}$$

This holds for both possible values of $l_{m'}$.

This result is valid for any f, and therefore there is no need to average explicitly over f. So to complete the calculation we just have to perform the average over the realizations of the training set. This is done in exactly the same way as for the BO case (Eqs.36,35), to give the following:

$$E(C \mid U = Circ, m', m, G) = 2 \binom{n}{m'}^{-1} \sum_{m_1=0}^{m'} \binom{R}{m_1} \binom{n - R}{m' - m_1} \frac{R - m_1}{n - m'} \left[1 - \frac{R - m_1}{n - m'}\right] . \tag{47}$$

The case of $m = 0$ is immediately retrieved as

$$E(C \mid U = Circ, m' = 0, m, G) = 2r(1 - r) \tag{48}$$

(note that $r = R/n < 1/2$). This is also the asymptotic result for $n \to \infty$ (obtained in the same way as in the BO case). The curves of $E(C \mid U = Circ, m', m, G)$ vs. m', for $n = 128$ and $R = 32, 60$, are also shown in Figure 1 with the BO results.

5 Discussion

In this paper we perform a preliminary investigation of learning curves for Bayes-optimal and Gibbs learning algorithms, for both IID and OTS error. In particular, we prove that for a uniform sampling distribution $\pi(x)$, for any loss function and any noise process, the learning curve for OTS error for a Bayes optimal generalizer is non-increasing. (It can increase for non-uniform $\pi(x)$.) We also characterize those priors for which the learning curve is constant. We also prove the similar result that for a uniform $\pi(x)$ the Gibbs generalizer has a non-increasing learning curve, provided certain (common) conditions on the loss function are met.

There are many questions that our analysis raises. Examples are:

i) What is the behavior of $E(C \mid m, G) - E(C \mid m, BO)$?

ii) What are the shapes (and in particular the widths) of the distributions whose means are examined above? ("Shapes" as one varies f, varies d, etc.) What governs quantities like $E((C_1 - C_2)^2 \mid m_1, m_2 > m_1)$? ($c_i$ is the error for the case of a training set of size m_i , and it's required that the training sets of size m_2 be compatible with (i.e., contain) the training sets of size m_1.)

iii) For non-uniform $\pi(x)$, is it possible for $E(C \mid m)$ to increase for one region of values of m and decrease for another, for a fixed $P(f)$?

iv) More generally, what characteristics of $\pi(x)$ and $P(f)$ determine $E(C \mid m)$'s behavior? (The analysis above provides a general overview of the relation between those variables, but hardly an exhaustive analysis.) In particular, how do these results change if we average over all π rather than fix π? What fraction of the set of all π's and $P(f)$'s result in regions of increasing expected OTS error?

v) One can define "Bayes-optimal" with respect to a class H: the guess made by the algorithm is the $h \in H$ that minimizes $E(C \mid d, h)$. Future work involves investigating the OTS behavior of such "partially optimal" algorithms. (See [9] for related previous work.)

vi) As in [7], one can imagine that one's learning algorithm is based on an incorrect guess for the prior over targets. The OTS behavior for such scenarios is currently unknown.

vii) Do the results given above relating ideals, constancy of learning curves, and variability over f hold for Gibbs generalizers as well as Bayes-optimal generalizers?

vii) We have a general outline of how things change when one conditions on m' rather than m (see above) but know little about the details. In particular, what happens when m and m' both vary and there is noise?

ix) How do things change if we allow $\pi(x)$ to be statistically coupled to f (as it almost always is in the real world)?

x) How do things change when we use other generalizers (besides the Bayes-optimal and Gibbs generalizers) that also can be viewed as making an assumption for the prior? Examples of such generalizers are non-zero-temperature Boltzmann distribution generalizers, Bayesian maximum a posterior estimators (and their close cousins, neural nets taught by back propagation—see [wolpert-nips93]), etc.

xii) One of the referees suggested that the results in [7] providing upper bounds on cumulative error for Gibbs and Bayes optimal generalizers in terms of the VC dimension can be extended to OTS error. This and other possible extensions of the results in [7] bear future investigation.

Acknowledgements We thank the referees for their helpful comments. DHW was supported by The Santa Fe Institute and TXN Inc. for financial support, and EK and TG were supported by the Department of Energy.

References

[1] Bayarri, M.J, and Berger, J.O., "Applications and limitations of robust Bayesian bounds and type II MLE", Purdue University Department of Statistics TR 93-11C.

[2] Berger, J. *Statistical Decision Theory and Bayesian Analysis.* New York: Springer-Verlag, 1985.

[3] Blumer, A., et al., "Occam's Razor", *Info. Proc. Lett.*, **24**, 377-380, 1987.

[4] Chaloner, K., and Verdinelli, I., "Bayesian experimental design: a review", University of Minnesota Minneapolis, Department of Statistics.

[5] Dietterich T. *Ann. Rev. Comp. Sci.* **4** (1990): 255–306.

[6] Cussens, J., "A Bayesian analysis of algorithms for learning finite functions", Glasgow Caledonian University, Department of Mathematics.

[7] Haussler D., Kearns M. and Schapire R., *Machine Learning*, **14**, pp. 83-115, 1994.

[8] Hill S.D. and Spall J.C., "Sensitivity of a bayesian Analysis to the Prior Distribution", *IEEE TSMC* **24**, 216-221, 1994

[9] Kearns, M. J., et al. "Towards efficient agnostic learning", in *Proceedings of the 5th annual workshop on Computational Learning Theory*, ACM Press, NY, NY, 1992.

[10] Quote from Machine Learning course taught by Tom Mitchell of Carnegie Mellon University Computer Science Department, fall of 1994.

[11] Gustafson, P., "The local Sensitivity of posterior expectations", Carnegie Mellon University, Department of Statistics.

[12] Opper M., and Haussler D., in *Proceedings of the 4th annual workshop on Computational Learning Theory*, 75–87. Morgan Kaufmann, 1991.

[13] Opper M. and Haussler D., *Phys. Rev. Lett.* **66** (1991) 2677-2680.

[14] Parzen, E., *Modern Probability Theory and its Applications*, Wiley, NY, NY, 1960.

[15] Schwartz D., Samalam V., Solla S., and Denker J., "Exhaustive Learning", *Neural Computation*, **2** (1990): 374-385.

[16] Tishby N., Levin E. and Solla S., in *International Joint Conference on Neural Networks, Vol. II*, (IEEE, 1989) 403–409.

[17] C. Van der Broeck and R. Kawai, *Phys. Rev. A.* **42** (1990) 6210-6218. and in the *Proc. of Intl. AMSE Conference on Neural Networks, San Diego (USA), May 1991, Vol.1, pp. 151-162.*

[18] Vapnik V., *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1992.

[19] Weiss S.M. and Kulikowski C.A. *Computer Systems that Learn.* San Mateo, CA: Morgan Kauffman, 1991.

[20] Wolpert D. In *Neural Information Processing Systems 6*, edited by S. Hanson et al. San Mateo, CA: Morgan-Kauffman, 1994.

[21] Wolpert D. and Lapedes A., in *The Mathematics of Generalization*, D. Wolpert Ed., 243-278, Addison-Wesley, 1994.

[22] Wolpert D. in *The Mathematics of Generalization*, D. Wolpert Ed., 117–214, Addison-Wesley, 1994.

[23] Wolpert D., *Off-Training Set Error and a priori distinctions between Learning Algorithms*, SFI TR 95-01-003.